Build Your Own Data Recipes

We already familiar with the basic structure and global arguments of data recipes or data configs of Data-Juicer. In this noteobok, we will learn how to build your own data recipes based on the existing recipes or the config_all.yaml file.

We simply set the basic global arguments only for input/output dataset paths and the number of subprocesses, so we can focus on the operator list refinement more.

We use the demo datasets as an example.

```
project_name: 'build_my_own_recipe'
dataset_path: '../demos/data/demo-dataset.jsonl'  # replace it
with the path to your dataset directory or file
np: 4  # number of subprocess to process your dataset
export_path: './outputs/my_own_recipe/res.jsonl'

process:
  - language_id_score_filter:
      lang: 'zh'
      min_score: 0.8
```

Now we will decide what kind of OPs we need to add.

Operator list

Data-Juicer offers an extensive array of OPs for data manipulation, encompassing modification, cleansing, filtering, and deduplication tasks.

Data recipe must include the necessary OPs and their respective arguments for efficient dataset processing. And Data-Juicer will process the OPs sequentially as arranged in the provided OP list.

Based on the OP list that contains only one OP above, we can add some other useful OPs.

For example, for textual samples, we can add a `whitespace_normalization_mapper` to normalize the whitespaces in the text to standard ASCII whitespace characters, which are more friendly to tokenizers of LLMs. Besides, deduplication is always necessary for some large-scale datasets to improve the training efficiency, so we can add a `document_deduplicator` to remove those duplicate texts from the dataset.

This would result the following data recipe and we can write it to a config file in YAML format:

In [1]:

```
config_str = """
project_name: 'build_my_own_recipe'
dataset_path: '../demos/data/demo-dataset.jsonl'  # replace it with the p
```

```
np: 4  # number of subprocess to process your dataset
export_path: './outputs/my_own_recipe/res.jsonl'

process:
  - whitespace_normalization_mapper:
  - language_id_score_filter:
      lang: 'zh'
      min_score: 0.8
  - document_deduplicator: # deduplicate text samples using md5 hashing e
      lowercase: false   # whether to convert text to lower case
      ignore_non_character: false
"""
recipe_name = 'my_own_recipe.yaml'
with open(recipe_name, 'w') as fout:
    fout.write(config_str)
```

Load and check the recipe

In [2]:

```
from data_juicer.config import init_configs
cfg = init_configs(args=f'--config {recipe_name}'.split())
print(f'np = {cfg.np}')
```

```
/usr/local/python310/lib/python3.10/site-packages/tqdm/auto.py:21: TqdmWar
ning: IProgress not found. Please update jupyter and ipywidgets. See http
s://ipywidgets.readthedocs.io/en/stable/user_install.html
  from .autonotebook import tqdm as notebook_tqdm
2024-08-08 12:17:04 | INFO     | data_juicer.config.config:618 - Back up t
he input config file [/root/projects/kdd_tutorial_notebooks/my_own_recipe.
yaml] into the work_dir [/root/projects/kdd_tutorial_notebooks/outputs/my_
own_recipe]
2024-08-08 12:17:04 | INFO     | data_juicer.config.config:640 - Configura
tion table:
╒═════════════════════════╤═══════════════════════════════════════════════
═══════════════════════════╕
│ key                      │ values
│
╞═════════════════════════╪═══════════════════════════════════════════════
═══════════════════════════╡
│ config                   │ [Path_fr(my_own_recipe.yaml, cwd=/root/project
s/kdd_tutorial_notebooks)] │
├─────────────────────────┼───────────────────────────────────────────────
───────────────────────────┤
│ hpo_config               │ None
│
├─────────────────────────┼───────────────────────────────────────────────
───────────────────────────┤
│ data_probe_algo          │ 'uniform'
│
├─────────────────────────┼───────────────────────────────────────────────
───────────────────────────┤
│ data_probe_ratio         │ 1.0
│
├─────────────────────────┼───────────────────────────────────────────────
───────────────────────────┤
│ project_name             │ 'build_my_own_recipe'
│
├─────────────────────────┼───────────────────────────────────────────────
───────────────────────────┤
│ executor_type            │ 'default'
│
├─────────────────────────┼───────────────────────────────────────────────
───────────────────────────┤
│ dataset_path             │ '/root/projects/data-juicer/demos/data/demo-da
taset.jsonl'               │
├─────────────────────────┼───────────────────────────────────────────────
───────────────────────────┤
│ export_path              │ '/root/projects/kdd_tutorial_notebooks/output
s/my_own_recipe/res.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                 │ [{'whitespace_normalization_mapper': {'acceler
ator': None,               │
│                         │                                       'audio_k
ey': 'audios',             │
│                         │                                       'cpu_req
uired': 1,                 │
│                         │                                       'image_k
ey': 'images',             │
│                         │                                       'mem_req
uired': 0,                 │
│                         │                                       'num_pro
c': 4,                     │
│                         │                                       'text_ke
y': 'text',                │
│                         │                                       'video_k
ey': 'videos'}},           │
│                         │  {'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'}},                  │
│                         │  {'document_deduplicator': {'accelerator': Non
e,                         │
│                         │                             'audio_key': 'audi
os',                       │
│                         │                             'cpu_required': 1,
│
│                         │                             'ignore_non_charac
ter': False,               │
│                         │                             'image_key': 'imag
es',                       │
│                         │                             'lowercase': Fals
e,                          │
│                         │                             'mem_required': 0,
│
│                         │                             'num_proc': 4,
│
│                         │                             'text_key': 'tex
t',                          │
│                         │                             'video_key': 'vide
os'}}]                     │
├─────────────────────────┼───────────────────────────────────────────────
───────────────────────────┤
│ 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/my_own_recipe'            │
├─────────────────────────┼───────────────────────────────────────────────
───────────────────────────┤
│ timestamp               │ '20240808121703'
│
├─────────────────────────┼───────────────────────────────────────────────
───────────────────────────┤
│ dataset_dir             │ '/root/projects/data-juicer/demos/data'
│
├─────────────────────────┼───────────────────────────────────────────────
───────────────────────────┤
│ add_suffix              │ False
│
╘═════════════════════════╧═══════════════════════════════════════════════
═══════════════════════════╛
np = 4
```

Now you can process the demo dataset with this new data recipe.

```
In [3]: !dj-process --config my_own_recipe.yaml
```

```
2024-08-08 12:17:24 | INFO     | data_juicer.config.config:618 - Back up t
he input config file [/root/projects/kdd_tutorial_notebooks/my_own_recipe.
yaml] into the work_dir [/root/projects/kdd_tutorial_notebooks/outputs/my_
own_recipe]
2024-08-08 12:17:24 | INFO     | data_juicer.config.config:640 - Configura
tion table:
╒═════════════════════════╤═══════════════════════════════════════════════
═══════════════════════════╕
│ key                      │ values
│
╞═════════════════════════╪═══════════════════════════════════════════════
═══════════════════════════╡
│ config                   │ [Path_fr(my_own_recipe.yaml, cwd=/root/project
s/kdd_tutorial_notebooks)] │
├─────────────────────────┼───────────────────────────────────────────────
───────────────────────────┤
│ hpo_config               │ None
│
├─────────────────────────┼───────────────────────────────────────────────
───────────────────────────┤
│ data_probe_algo          │ 'uniform'
│
├─────────────────────────┼───────────────────────────────────────────────
───────────────────────────┤
│ data_probe_ratio         │ 1.0
│
├─────────────────────────┼───────────────────────────────────────────────
───────────────────────────┤
│ project_name             │ 'build_my_own_recipe'
│
├─────────────────────────┼───────────────────────────────────────────────
───────────────────────────┤
│ executor_type            │ 'default'
│
├─────────────────────────┼───────────────────────────────────────────────
───────────────────────────┤
│ dataset_path             │ '/root/projects/data-juicer/demos/data/demo-da
taset.jsonl'               │
├─────────────────────────┼───────────────────────────────────────────────
───────────────────────────┤
│ export_path              │ '/root/projects/kdd_tutorial_notebooks/output
s/my_own_recipe/res.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                 │ [{'whitespace_normalization_mapper': {'acceler
ator': None,               │
│                         │                                       'audio_k
ey': 'audios',             │
│                         │                                       'cpu_req
uired': 1,                 │
│                         │                                       'image_k
ey': 'images',             │
│                         │                                       'mem_req
uired': 0,                 │
│                         │                                       'num_pro
c': 4,                     │
│                         │                                       'text_ke
y': 'text',                │
│                         │                                       'video_k
ey': 'videos'}},           │
│                         │  {'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'}},                  │
│                         │  {'document_deduplicator': {'accelerator': Non
e,                         │
│                         │                             'audio_key': 'audi
os',                       │
│                         │                             'cpu_required': 1,
│
│                         │                             'ignore_non_charac
ter': False,               │
│                         │                             'image_key': 'imag
es',                       │
│                         │                             'lowercase': Fals
e,                          │
│                         │                             'mem_required': 0,
│
│                         │                             'num_proc': 4,
│
│                         │                             'text_key': 'tex
t',                          │
│                         │                             'video_key': 'vide
os'}}]                     │
├─────────────────────────┼───────────────────────────────────────────────
───────────────────────────┤
│ 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/my_own_recipe'            │
├─────────────────────────┼───────────────────────────────────────────────
───────────────────────────┤
│ timestamp               │ '20240808121724'
│
├─────────────────────────┼───────────────────────────────────────────────
───────────────────────────┤
│ dataset_dir             │ '/root/projects/data-juicer/demos/data'
│
├─────────────────────────┼───────────────────────────────────────────────
───────────────────────────┤
│ add_suffix              │ False
│
╘═════════════════════════╧═══════════════════════════════════════════════
═══════════════════════════╛
2024-08-08 12:17:24 | INFO     | data_juicer.core.executor:47 - Using cach
e compression method: [None]
2024-08-08 12:17:24 | INFO     | data_juicer.core.executor:52 - Setting up
data formatter...
2024-08-08 12:17:24 | INFO     | data_juicer.core.executor:74 - Preparing
exporter...
2024-08-08 12:17:24 | INFO     | data_juicer.core.executor:151 - Loading d
ataset from data formatter...
```

```
2024-08-08 12:17:25 | INFO     | data_juicer.format.formatter:185 - Unifyi
ng the input dataset formats...
2024-08-08 12:17:25 | INFO     | data_juicer.format.formatter:200 - There
are 6 sample(s) in the original dataset.
2024-08-08 12:17:25 | INFO     | data_juicer.format.formatter:214 - 6 samp
les left after filtering empty text.
2024-08-08 12:17:25 | INFO     | data_juicer.format.mixture_formatter:137
- sampled 6 from 6
2024-08-08 12:17:25 | INFO     | data_juicer.format.mixture_formatter:143
- There are 6 in final dataset
2024-08-08 12:17:25 | INFO     | data_juicer.core.executor:157 - Preparing
process operators...
2024-08-08 12:17:25 | 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.
2024-08-08 12:17:25 | INFO     | data_juicer.core.executor:164 - Processin
g data...
2024-08-08 12:17:25 | INFO     | data_juicer.core.data:193 - OP [whitespac
e_normalization_mapper] Done in 0.010s. Left 6 samples.
2024-08-08 12:17:25 | INFO     | data_juicer.core.data:193 - OP [language_
id_score_filter] Done in 0.027s. Left 2 samples.
num_proc must be <= 2. Reducing num_proc to 2 for dataset of size 2.
2024-08-08 12:17:25 | INFO     | data_juicer.core.data:193 - OP [document_
deduplicator] Done in 0.009s. Left 2 samples.
2024-08-08 12:17:25 | INFO     | data_juicer.core.executor:171 - All OPs a
re done in 0.047s.
2024-08-08 12:17:25 | INFO     | data_juicer.core.executor:174 - Exporting
dataset to disk...
2024-08-08 12:17:25 | 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, 239.80
ba/s]
2024-08-08 12:17:25 | 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, 1136.3
6ba/s]
```

Finally we clean up the temporary recipe.

In [4]:
```
!rm my_own_recipe.yaml
```

Build Method

In addition to modifying from existing built-in recipes, you can also:

- ## Customize the Default Configuration File

The `config_all.yaml` contains all operators and their default arguments.

You just need to **remove** ops that you won't use and refine some arguments of ops.

- ## Create a New Configuration from Scratch

You can refer our example config file `config_all.yaml`, op documents, and advanced Build-Up Guide for developers and create a new recipe from scratch.

Reuseble Built-in Recipes

Data-Juice offers tens of built-in data processing recipes for pre-training, fine-tuning, en, zh, and more scenarios.

Reproduced Redpajama

We have reproduced the processing flow of some RedPajama datasets. Please refer to the reproduced_redpajama folder for details.

Reproduced BLOOM

We have reproduced the processing flow of some BLOOM datasets. please refer to the reproduced_bloom folder for details.

Data-Juicer Recipes

We have refined some open source datasets (including CFT datasets) by using Data-Juicer and have provided configuration files for the refined flow. please refer to the data_juicer_recipes folder for details.

Awesome LLM Data

We provide a tag-based categorization to help readers easy diving into the myriad of materials, promoting an intuitive understanding of each entry's key focus areas. Soon we will provide a dynamic table of contents to help readers more easily navigate through the materials with features such as search, filter, and sort.

For more detail, please refer to Awesome LLM Data

Conclusion

In this notebook, we learn how to build our own recipes from existing recipes. And we show that Data-Juicer already prepared lots of built-in data recipes for users to refer.