Analyze & Probe Datasets

In this notebook, we will introduce how to analyze and probe given datasets using Data-Juicer Analyzer tools. By using the Analyzer, we can obtain statistical information about the dataset and use these statistics to set and refine parameters in the data recipes.

Note: Analyzer only computes and analyze the stats of Filter operators. It doesn't work for Mappers, Deduplicators, and Selectors.

Similar to data processing, we can run `analyze_data.py` tool or `dj-analyze` command with your config as the argument to analyze your dataset. Both of them use `Analyzer` to finish the analysis.

```
# only for installation from source
python tools/analyze_data.py --config your_recipe.yaml

# use command line tool
dj-analyze --config your_recipe.yaml
```

Here, we will show you how to analyze your dataset.

We will also use the demo dataset in Data-Juicer in this example.

First, we need to prepare a data recipe for analyzing the dataset, which includes Filters with stats we care about.

For example, if we want to check the distribution of text length of this dataset, we only need to add `text_length_filter` OP in the OP list. Because we won't filter out any samples in the dataset, we don't need to set the thresholds of this OP. And for convenience, we save all analysis results into one figure file.

In [1]:
```
recipe_str = """
project_name: 'analyze_a_dataset'
dataset_path: '../demos/data/demo-dataset.jsonl'  # path to your dataset
np: 1  # number of subprocess to process your dataset

export_path: './outputs/analyze_result/res.jsonl'
save_stats_in_one_file: true

# process schedule
# a list of several process operators with their arguments
process:
  - text_length_filter:                                    # filter text
"""

analyze_recipe = 'analyze_recipe.yaml'
with open(analyze_recipe, 'w') as f:
    f.write(recipe_str)
```

Now let's begin to analyze the dataset with `Analyzer` directly.

In [2]:

```
# load recipe
from data_juicer.config import init_configs
from data_juicer.core import Analyzer

cfg = init_configs(args=f'--config {analyze_recipe}'.split())

analyzer = Analyzer(cfg)
analyzer.run()
```

```
/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 17:18:55 | INFO     | data_juicer.config.config:618 - Back up t
he input config file [/root/projects/kdd_tutorial_notebooks/analyze_recip
e.yaml] into the work_dir [/root/projects/kdd_tutorial_notebooks/outputs/a
nalyze_result]
2024-08-08 17:18:55 | INFO     | data_juicer.config.config:640 - Configura
tion table:
╒═══════════════════════════╤═══════════════════════════════════════════════
═══════════════════════════╕
│ key                       │ values
│
╞═══════════════════════════╪═══════════════════════════════════════════════
═══════════════════════════╡
│ config                    │ [Path_fr(analyze_recipe.yaml, cwd=/root/projec
ts/kdd_tutorial_notebooks)] │
├───────────────────────────┼───────────────────────────────────────────────
──────────────────────────┤
│ hpo_config                │ None
│
├───────────────────────────┼───────────────────────────────────────────────
──────────────────────────┤
│ data_probe_algo           │ 'uniform'
│
├───────────────────────────┼───────────────────────────────────────────────
──────────────────────────┤
│ data_probe_ratio          │ 1.0
│
├───────────────────────────┼───────────────────────────────────────────────
──────────────────────────┤
│ project_name              │ 'analyze_a_dataset'
│
├───────────────────────────┼───────────────────────────────────────────────
──────────────────────────┤
│ executor_type             │ 'default'
│
├───────────────────────────┼───────────────────────────────────────────────
──────────────────────────┤
│ dataset_path              │ '/root/projects/data-juicer/demos/data/demo-da
taset.jsonl'                │
├───────────────────────────┼───────────────────────────────────────────────
──────────────────────────┤
│ export_path               │ '/root/projects/kdd_tutorial_notebooks/output
s/analyze_result/res.jsonl'  │
├───────────────────────────┼───────────────────────────────────────────────
──────────────────────────┤
│ export_shard_size         │ 0
│
├───────────────────────────┼───────────────────────────────────────────────
──────────────────────────┤
│ export_in_parallel        │ False
│
├───────────────────────────┼───────────────────────────────────────────────
──────────────────────────┤
│ keep_stats_in_res_ds      │ False
│
├───────────────────────────┼───────────────────────────────────────────────
──────────────────────────┤
```

keep_hashes_in_res_ds	False
np	1
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                 | [{'text_length_filter': {'accelerator': None,
|                         |                          'audio_key': 'audio
s',                        |
|                         |                          'cpu_required': 1,
|                         |                          'image_key': 'image
s',                        |
|                         |                          'max_len': 9223372036
854775807,                 |
|                         |                          'mem_required': 0,
|                         |                          'min_len': 10,
|                         |                          'num_proc': 1,
|                         |                          'stats_export_path':
None,                      |
|                         |                          'text_key': 'text',
|                         |                          'video_key': 'video
s'}}]                      |
├─────────────────────────┼──────────────────────────────────────────────────────────────────────
──────────────────────────┤
| percentiles             | []
|
├─────────────────────────┼──────────────────────────────────────────────────────────────────────
──────────────────────────┤
| export_original_dataset | False
|
├─────────────────────────┼──────────────────────────────────────────────────────────────────────
──────────────────────────┤
| save_stats_in_one_file  | True
|
├─────────────────────────┼──────────────────────────────────────────────────────────────────────
──────────────────────────┤
| ray_address             | 'auto'
|
├─────────────────────────┼──────────────────────────────────────────────────────────────────────
──────────────────────────┤
```

```
| debug                    | False
|
├──────────────────────────┼──────────────────────────────────────────────────────────────────────
──────────────────────────┤
| work_dir                 | '/root/projects/kdd_tutorial_notebooks/output
s/analyze_result'          |
├──────────────────────────┼──────────────────────────────────────────────────────────────────────
──────────────────────────┤
| timestamp                | '20240808171854'
|
├──────────────────────────┼──────────────────────────────────────────────────────────────────────
──────────────────────────┤
| dataset_dir              | '/root/projects/data-juicer/demos/data'
|
├──────────────────────────┼──────────────────────────────────────────────────────────────────────
──────────────────────────┤
| add_suffix               | False
|
╘══════════════════════════╧══════════════════════════════════════════════════════════════════════
══════════════════════════╛
2024-08-08 17:18:55 | INFO     | data_juicer.core.analyzer:37 - Using cach
e compression method: [None]
2024-08-08 17:18:55 | INFO     | data_juicer.core.analyzer:42 - Setting up
data formatter...
2024-08-08 17:18:55 | INFO     | data_juicer.core.analyzer:51 - Preparing
exporter...
2024-08-08 17:18:55 | INFO     | data_juicer.core.analyzer:75 - Loading da
taset from data formatter...
2024-08-08 17:18:56 | INFO     | data_juicer.format.formatter:185 - Unifyi
ng the input dataset formats...
2024-08-08 17:18:56 | INFO     | data_juicer.format.formatter:200 - There
are 6 sample(s) in the original dataset.
2024-08-08 17:18:56 | INFO     | data_juicer.format.formatter:214 - 6 samp
les left after filtering empty text.
2024-08-08 17:18:56 | INFO     | data_juicer.format.mixture_formatter:137
- sampled 6 from 6
2024-08-08 17:18:56 | INFO     | data_juicer.format.mixture_formatter:143
- There are 6 in final dataset
2024-08-08 17:18:56 | INFO     | data_juicer.core.analyzer:81 - Preparing
process operators...
2024-08-08 17:18:56 | INFO     | data_juicer.core.analyzer:86 - Computing
the stats of dataset...
2024-08-08 17:18:56 | INFO     | data_juicer.core.data:193 - OP [text_leng
th_filter] Done in 0.007s. Left 6 samples.
2024-08-08 17:18:56 | INFO     | data_juicer.core.analyzer:101 - Exporting
dataset to disk...
2024-08-08 17:18:56 | 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, 230.58
ba/s]
2024-08-08 17:18:56 | INFO     | data_juicer.core.analyzer:113 - Applying
overall analysis on stats...
100%|##########| 1/1 [00:00<00:00, 8594.89it/s]
2024-08-08 17:18:56 | INFO     | data_juicer.core.analyzer:120 - The overa
ll analysis results are:          text_len
count    6.000000
mean    39.500000
std     34.795115
min      8.000000
25%     13.750000
```

```
50%     32.500000
75%     49.750000
max    101.000000
2024-08-08 17:18:56 | INFO     | data_juicer.core.analyzer:122 - Applying column-wise analysis on stats...
Column: 100%|##########| 1/1 [00:00<00:00, 48.18it/s]
```

Out[2]:
```
Dataset({
    features: ['text', 'meta', '__dj__stats__'],
    num_rows: 6
})
<Figure size 800x600 with 0 Axes>
```

After the analysis is complete, we can view the statistical information of the dataset.

In [3]:
```
import os
import pandas as pd
overall_file = os.path.join(analyzer.analysis_path, 'overall.csv')
if os.path.exists(overall_file):
    analysis_res = pd.read_csv(overall_file)

analysis_res
```

Out[3]:

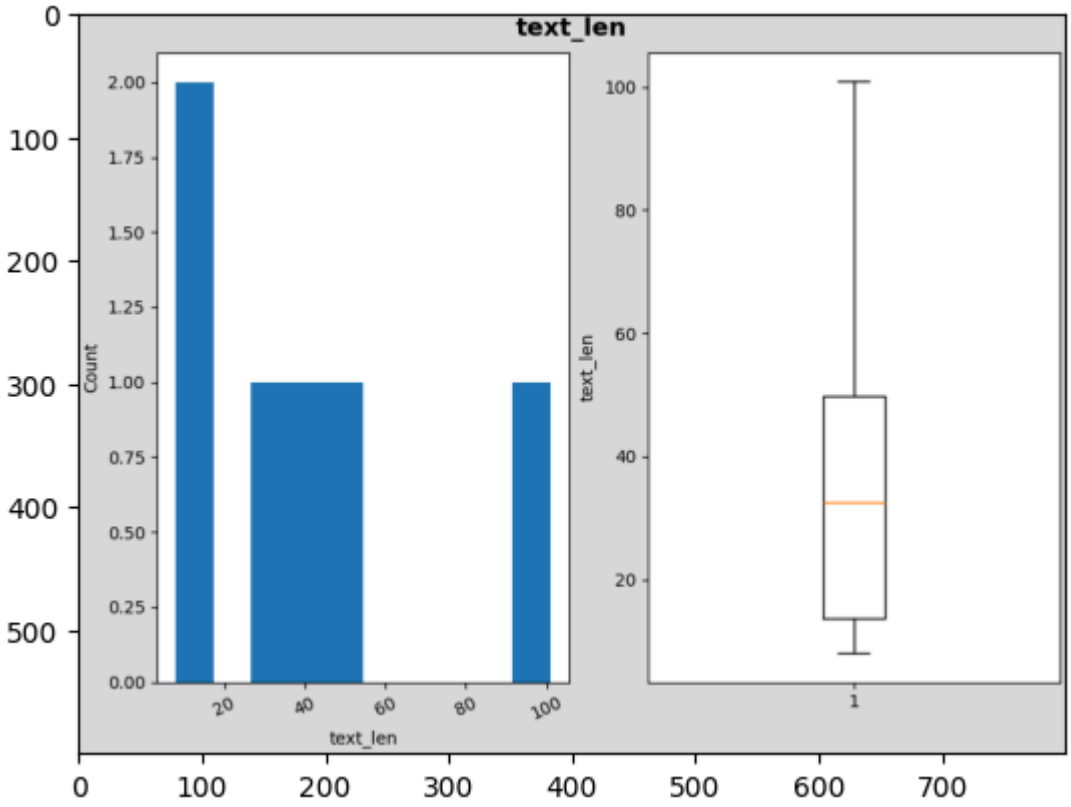
	Unnamed: 0	text_len
0	count	6.000000
1	mean	39.500000
2	std	34.795115
3	min	8.000000
4	25%	13.750000
5	50%	32.500000
6	75%	49.750000
7	max	101.000000

Display the histogram of statistics of dataset

In [4]:
```
import matplotlib.pyplot as plt
import matplotlib.image as mpimg

if os.path.exists(analyzer.analysis_path):
    for f_path in os.listdir(analyzer.analysis_path):
        if '.png' in f_path and 'all-stats' in f_path:
            all_stats = os.path.join(analyzer.analysis_path, f_path)
            break

img = mpimg.imread(all_stats)
plt.imshow(img)
plt.show()
```



Finally, clean up the example recipe.

In [5]:
```
!rm analyze_recipe.yaml
```

Conclusion

In this notebook, we learn how to analyze a dataset on Filter stats we care about with the `Analyzer` tool, and how to check the analysis results.