

Thai Rule Set Tech stack

- FastAPI (Service as API)
- pyThaiNLP (Data Discovery tool for Thai language)
- Dask-Dataframes (Parallel processing)
- Faust (Streaming processing)

Demo

<https://docs.dask.org/en/latest/remote-data-services.html#optional-parameters>

<http://127.0.0.1:8000/docs>

<http://localhost:8787/status>

https://colab.research.google.com/drive/1P-jk4eyzPuckSMaH2eb1DN16Ec_dUBkRL

DataToolsAPI 0.1 OAS3

[/openapi.json](#)

A REST API for data discovery and data refinery in Thai

default



POST

/api/0.1/data_discovery/text Data Discovery Text

POST

/api/0.1/data_discovery/text_list Data Discovery Text List

POST

/api/0.1/data_discovery/text_file Data Discovery Text File

POST

/api/0.1/data_refinery/text Data Refinery Text

POST

/api/0.1/data_refinery/text_list Data Refinery Text List

POST

/api/0.1/data_refinery/text_file Data Refinery Text File

Features of DataToolsAPI

- Data Discovery (pyThaiNLP)
- Data Refinery (Thai Rule Set 5 modules)

The *B*- prefix indicates beginning token for a chunk of person name, “มาร์ค โอบามา” and *I*- prefix indicates the intermediate token. However, the term *O* indicates that a token not belong to any NER chunk.

The following table shows the list of Named Entity Recognition (NER) tags:

Named Entity Recognition tag	Examples
DATE	2/21/2004, 16 ก.พ., จันทร์
TIME	16.30 น., 5 วัน, 1-3 ปี
EMAIL	info@nrpsc.ac.th
LEN	30 กิโลเมตร, 5 กม.
LOCATION	ไทย, จ.ปราจีนบุรี, กำแพงเพชร
ORGANIZATION	กรมวิทยาศาสตร์การแพทย์, อย.
PERSON	น.พ.จรัส, นางประนอม ทองจันทร์
PHONE	1200, 0 2670 8888
URL	http://www.bangkokhealth.com/
ZIP	10400, 11130
Money	2.7 ล้านบาท, 2,000 บาท
LAW	พ.ร.บ.โรคระบาด พ.ศ.2499, รัฐธรรมนูญ

<https://www.thainlp.org/pythainlp/docs/2.0/api/tag.html#>

Meet minimum requirements

- Implement the first version Thai Rule Set as an API
- Can data discovery

Future work

- Real-time process streaming data with streamz or faust
- Implement module ที่เหลือของ QualityStage ไปยัง Python