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-jk4eyzPuckSMaH2eblDNl6EcdUBkRL

Core Features of DataToolsAPI

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

Input: text, text array, text file

The *B-* prefix indicates begining token for a chunk of person name, “บารัค 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.จรัล, 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 (**Naming only**)
- Can data discovery (PII)

Future work

- To make a Real-time process streaming data with streamz or faust (optional)
- To understand Pattern Action Reference of IBM QualityStage
- To implement remaining modules of IBM QualityStage into Python (4 models) ที่อยู่, ชื่อ, Email, Phone เรียงลำดับตามความยาก

Original Thai Rule Set version (IBM QualityStage) จากพี่เมย์

- Input ของลูกค้าเป็น Database (tables) หรือ File
- Output ให้ลูกค้าเป็น Database (tables) หรือ export File
- เรื่อง Location -> ทำ Similarity ได้
- ทำ custom dictionary ได้
- ทำ Data Discovery ได้ (optional)
  - isThaiAddress
  - isPersonalName
  - isPhone
  - isEmail
- Due date: Q2