PyThaiNLP: Thai Natural Language Processing in Python

A free and open-source natural language processing (NLP) library for Thai language.

Features in PyThaiNLP

Tokenizers
Character Cluster and Syllable Level
Word Level
Sentence Level

Phonetic Algorithm and Transliteration
Grapheme-to-Phoneme
Soundex
Thai-English Transliteration

Embedding
Word Level
Sentence Level

Sequence Tagging
Named-Entity Recognition
Part-of-Speech Tagging

Automatic Speech Recognition*

Co-reference and Entity Linking

Spell Checking

Machine Translation*

Datasets

VISTEC-TPTH-2020 (Limkonchotiwat et al., 2021)
Task: *Word Tokenization*; Domain: *social media*

SCB-MT-EN-TH* (Lowphansirikul et al., 2020)
Task: *Coreference Resolution*; Domain: *news and Wikipedia articles*

Thai NER (Phatthiyaphaibun, 2022)
Task: *Named-Entity Recognition*; Domain: *news and Wikipedia articles*

Han-Coref (Phatthiyaphaibun and Limkonchotiwat, 2023)
Task: *Coreference Resolution*; Domain: *news and Wikipedia articles*

Pre-trained Langauge Models

WangchanBERTa* (Lowphansirikul et al., 2021a)
Thai Pre-trained Language Model

WangchanGLM (Polpanumas et al., 2023)
Multilingual Instruction-Following Model

*: in collaboration with the VISTEC-depa Thailand Artificial Intelligence Research Institute

EMNLP 2023 PyThaiNLP/PyThaiNLP