

Indic NLP Library

A unified approach to NLP for Indian languages

Anoop Kunchukuttan (anoop.kunchukuttan@gmail.com)

The goal of the Indic NLP Library is to build Python based libraries for common text processing and Natural Language Processing in Indian languages. Indian languages share a lot of similarity in terms of script, phonology, language syntax, etc. and this library is an attempt to provide a general solution to very commonly required toolsets for Indian language text.

The library provides the following functionalities:

- Text Normalization
- Script Information
- Word Tokenization and Detokenization
- Sentence Splitting
- Word Segmentation
- Syllabification
- Script Conversion
- Romanization
- Indicization
- Transliteration
- Translation

The data resources required by the Indic NLP Library are hosted in a different repository. These resources are required for some modules. You can download from the [Indic NLP Resources](#) project.

If you are interested in Indian language NLP resources, you should check the [Indic NLP Catalog](#) for pointers.

Pre-requisites

- Python 3.x
 - (For Python 2.x version check the tag `PYTHON_2.7_FINAL_JAN_2019` . Not actively supporting Python 2.x anymore, but will try to maintain as much compatibility as possible)
- [Indic NLP Resources](#)
- Other dependencies are listed in setup.py

Configuration

- Installation from pip:

```
pip install indic-nlp-library
```

- If you want to use the project from the github repo, add the project to the Python Path:

- Clone this repository
 - Install dependencies: `pip install -r requirements.txt`
 - Run: `export PYTHONPATH=$PYTHONPATH:<project base directory>`
- In either case, export the path to the *Indic NLP Resources* directory

Run: `export INDIC_RESOURCES_PATH=<path to Indic NLP resources>`

Usage

You can use the Python API to access all the features of the library. Many of the most common operations are also accessible via a unified commandline API.

Getting Started

Check [this IPython Notebook](#) for examples to use the Python API. - You can find the Python 2.x Notebook [here](#)

Documentation

You can find detailed documentation [HERE](#)

This documents the Python API as well as the commandline reference.

Citing

If you use this library, please include the following citation:

```
@unpublished{kunchukuttan2020indicnlp,  
author = "Anoop Kunchukuttan",  
title = "The IndicNLP Library",  
year = "2020",  
}
```

You can find the document [HERE](#)

Website

http://anoopkunchukuttan.github.io/indic_nlp_library

Author

Anoop Kunchukuttan (anoop.kunchukuttan@gmail.com)

Version: 0.7

Revision Log

0.7 : 02 Apr 2020:

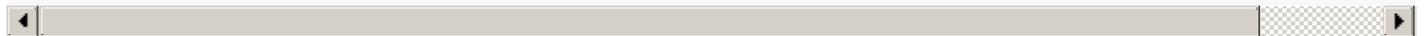
- Unified commandline
- Improved documentation
- Added setup.py

0.6 : 16 Dec 2019:

- New romanizer and indicizer
- Script Unifiers
- Improved script normalizers
- Added contrib directory for sample uses
- changed to MIT license

0.5 : 03 Jun 2019:

- Improved word tokenizer to handle dates and numbers.
- Added sentence splitter that can handle common prefixes/honorofics and uses some heuristics
- Added detokenizer
- Added acronym transliterator that can convert English acronyms to Brahmi-derived script



0.4 : 28 Jan 2019: Ported to Python 3, and lots of feature additions since last release; primarily around script information, script similarity and syllabification.

0.3 : 21 Oct 2014: Supports morph-analysis between Indian languages

0.2 : 13 Jun 2014: Supports transliteration between Indian languages and tokenization of Indian languages

0.1 : 12 Mar 2014: Initial version. Supports text normalization.

LICENSE

Indic NLP Library is released under the MIT license