Iranlowo

Release 0.1

Jul 06, 2019

Contents:

Ìrànlw is a set of utilities to analyze & process Yorùbá text for NLP tasks. The focus is on *helping software developers* build large, clean text datasets for (further) diacritic restoration and machine translation tasks.

CHAPTER 1

Features

1.1 ADR tools

- [X] Strip all diacritics from word-types
- [X] Verify that text is NFC or NFD
- [X] Canonicalize a corpus (from MS Word or elsewhere) → NFC
- [X] Split long sentences on certain characters like `;`, `:`, etc
- [X] Automatically restore correct diacritics using a pre-trained model
- [X] Find all variants of all word-type in a given corpus
- [ ] Partially strip diacritics from word-types

1.2 Ready to use webpage scrapers

- [X] Bíbélì Mím
- [X] Yoruba Bible - Bible Society of Nigeria
- [ ] Yorùbá Blog
- [ ] BBC Yorùbá

1.3 Corpus analysis tools

- [X] Dataset character distribution
- [X] Dataset ambuiguity statistics → Lexdif, etc for a given corpus
- [ ] Dataset scoring (proximity to correctly diacritized text, LM perplexity, KL divergence)

CHAPTER 2

Installation

Obtainable from the Python Package Index (PyPI) → `pip install iranlowo`

CHAPTER 3

Example

- Show computing environment and installation process
- Diacritize a phrase
- Diacritize phrases, note we use `ipython` only because it renders nicer, easy-to-read text-colours in the terminal!

CHAPTER 4

Disclaimer

This is beta software, if you pass the diacritizer out-of-domain text, English, pidgin or any other non-Yorùbá text, you will experience very marvelous, black-box results.

Since this a work-in-progress and we are steadily improving, if you encounter any problems with correctness or performance, please submit pull-requests with corrections or file an issue.

CHAPTER 5

License

This project is licensed under the MIT License.