

Data Collection Pipeline for Low-Resource Languages: A Case Study on Constructing a **Tetun** Text Corpus

Gabriel de Jesus, Sérgio Nunes

INESC TEC / Faculty of Engineering, University of Porto (FEUP)

The 2024 Joint International Conference on Computational Linguistics,
Language Resources, and Evaluation

Lingotto Conference Centre, Turin, Italy
20 - 25 May, 2024

Context and Motivation

- Text corpora are crucial for the advancement of IR and NLP tools.
- LRLs are characterized by data scarcity and linguistic complexities.
- Text corpora for LRLs are often unavailable.
- Interests in developing tools for LRLs have consistently risen.

Context and Motivation

Tetun

- The most widely spoken language in Timor-Leste.
- One of Timor-Leste's Official languages alongside Portuguese

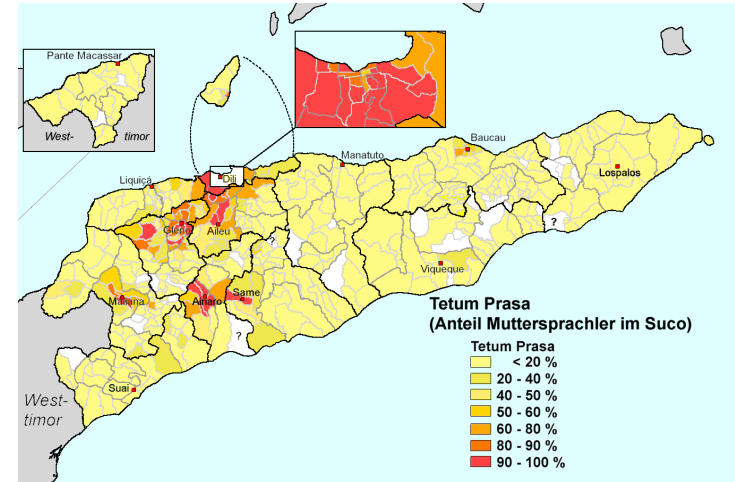
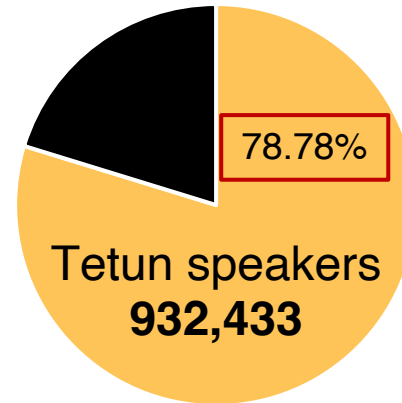


Image source: Wikipedia



1.18 million population

Context and Motivation

- A readily **Tetun corpus** is not available.
- Tetun **language processing components** do not exist.



A common technique for developing text corpora involves **crawling the web**.

Related Work

Artetxe et al. (2022):

Tailored crawling: Manually identifying data sources containing high-quality data and then scraping their contents to construct a text corpus for LRLs.

Implementation:

Constructing a Basque text corpus.

Challenges:

Data sources containing high-quality data.

Related Work

Korner et al. (2022):

Crawling and Collaborative:

Corpus creation through a web portal, which enabling community participations.

Implementation:

Constructing corpora for 258 LRLs.

Challenges:

Motivating community to participate.

Related Work

Wenzek et al. (2020):

Processing one Common Crawl snapshot to construct corpora, including for LRLs.

Implementation:

Constructing corpora for 174 languages, including LRLs such as Basque and Malay.

Challenges:

Require adequate computational power to process the CC snapshots.

Propose Solutions



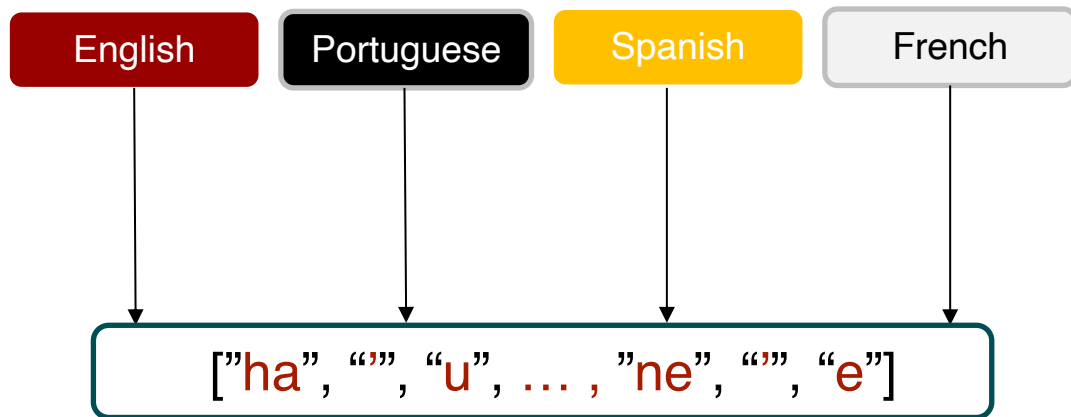
Labadain Crawler - A data collection pipeline that relies on three key components:

- An **initial text** in the target language.
- A **tokenizer**.
- A **language identification (LID)** model.

Tetun Tokenizer

Tetun input text: “ha’u-nia uma mak ne’e” (this is my home).

Applying the existing tokenizers in NLTK:



Tetun tokenizer:

```
pip install tetun-tokenizer
```

```
from tetuntokenizer.tokenizer  
import TetunSimpleTokenizer
```

["ha'u", ..., "ne'e"]

Tetun Tokenizer

Rule-based techniques using regular expressions:

- **Word tokenizer:** extracting only word units, **excluding** numbers, punctuation, and special characters.
- **Simple tokenizer:** extracting only words and numbers, **excluding** punctuation and special characters.

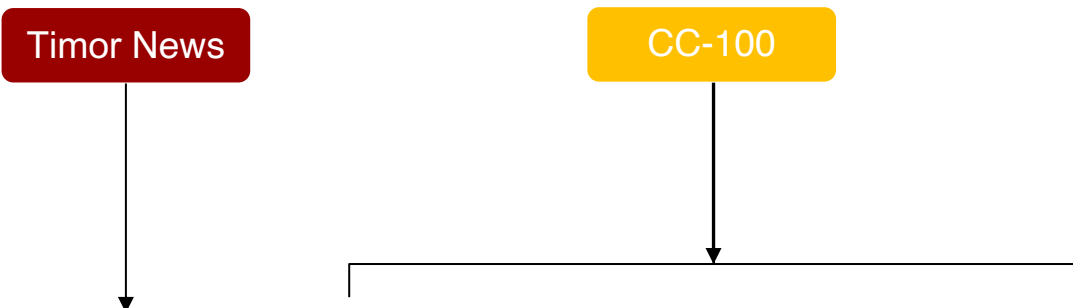
Tetun Tokenizer

Evaluations:

- Five Timorese volunteer students evaluated each tokenization techniques.
- Each student evaluated a minimum of three text samples (200 to 250 words) that were collected from the web.
- The evaluators reported that all the tokenizer techniques achieved 100% of accuracy for each tested input text.

Tetun Language Identification

Dataset



	Tetun	Portuguese	English	Indonesian
Total sentences	18,108	29,056	34,509	31,888
Minimum words per sentence	2	1	1	1
Maximum words per sentence	209	1,122	220	1,746
Average words per sentence	29.12	20.89	18.99	16.47
Total words in document	527,258	606,867	655,328	525,298

Tetun Language Identification

Training and evaluation

Input: sentences

Labels: languages

Training set: 70% Dev. set: 15%

Test set: 15%

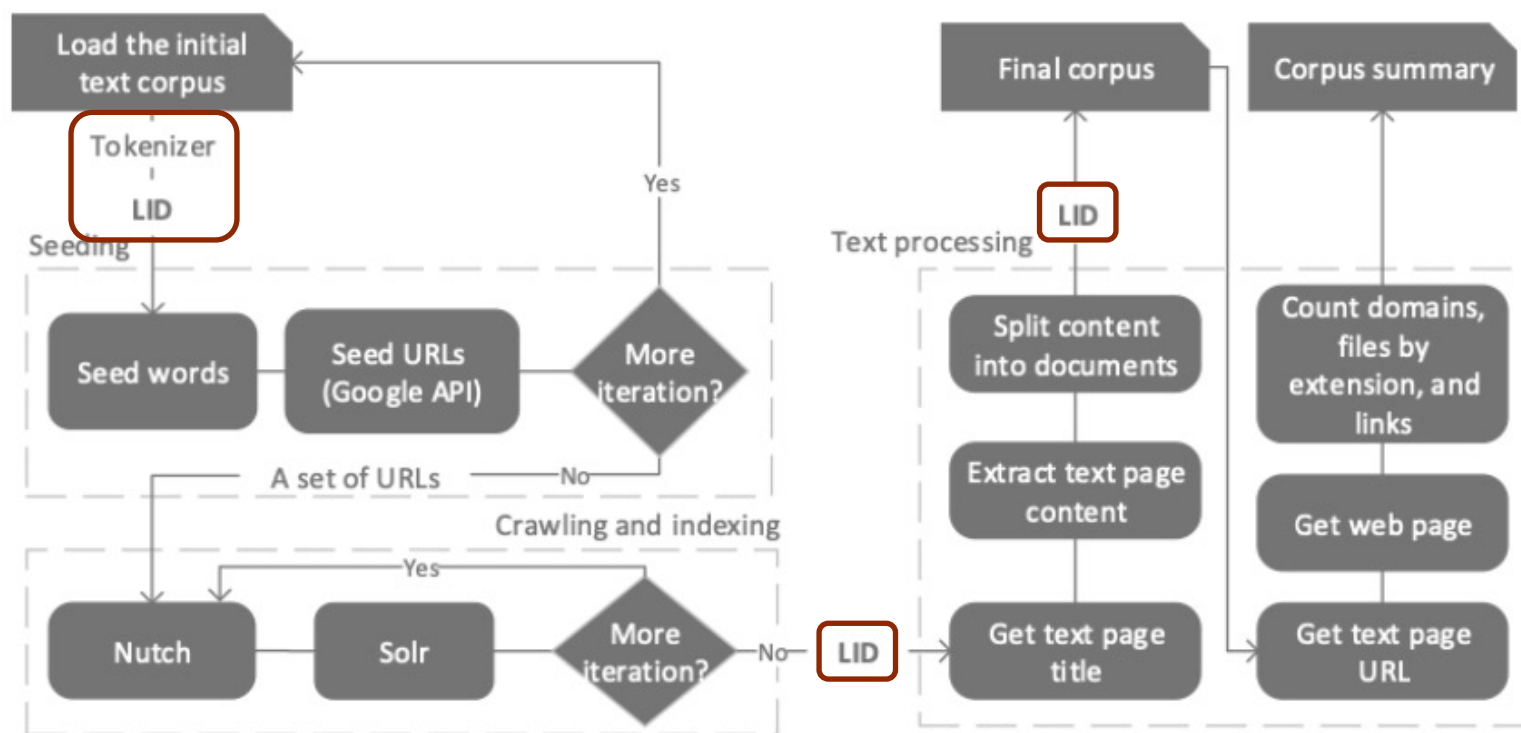
Model	Character n-gram						Word n-gram		
	1	2	3	4	5	6	1	2	3
SVM	0.9822	0.9954	0.9974	0.9975	0.9974	0.9968	0.9953	0.9689	0.8397
LR	0.9812	0.9953	0.9970	0.9975	0.9971	0.9960	0.9930	0.9522	0.8107
MNB	0.9452	0.9918	0.9967	0.9977	0.9981	0.9979	0.9973	0.9755	0.7806

Accuracy of the models' performance when evaluating using the development set.

	Overall Accuracy	F1
Tetun	0.9977	0.9987
Portuguese		0.9984
English		0.9976
Indonesia		0.9979

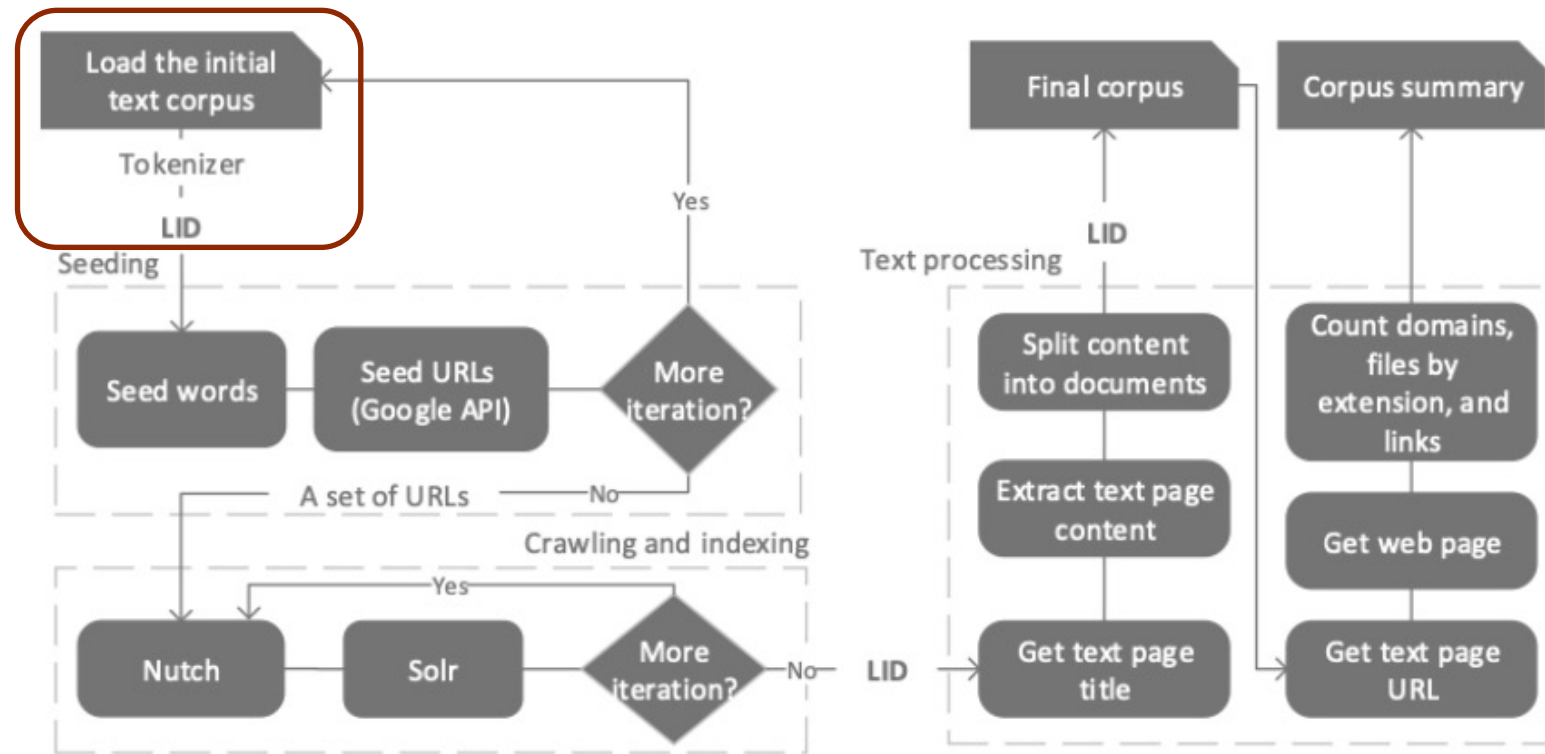
Labadain Crawler

General architecture



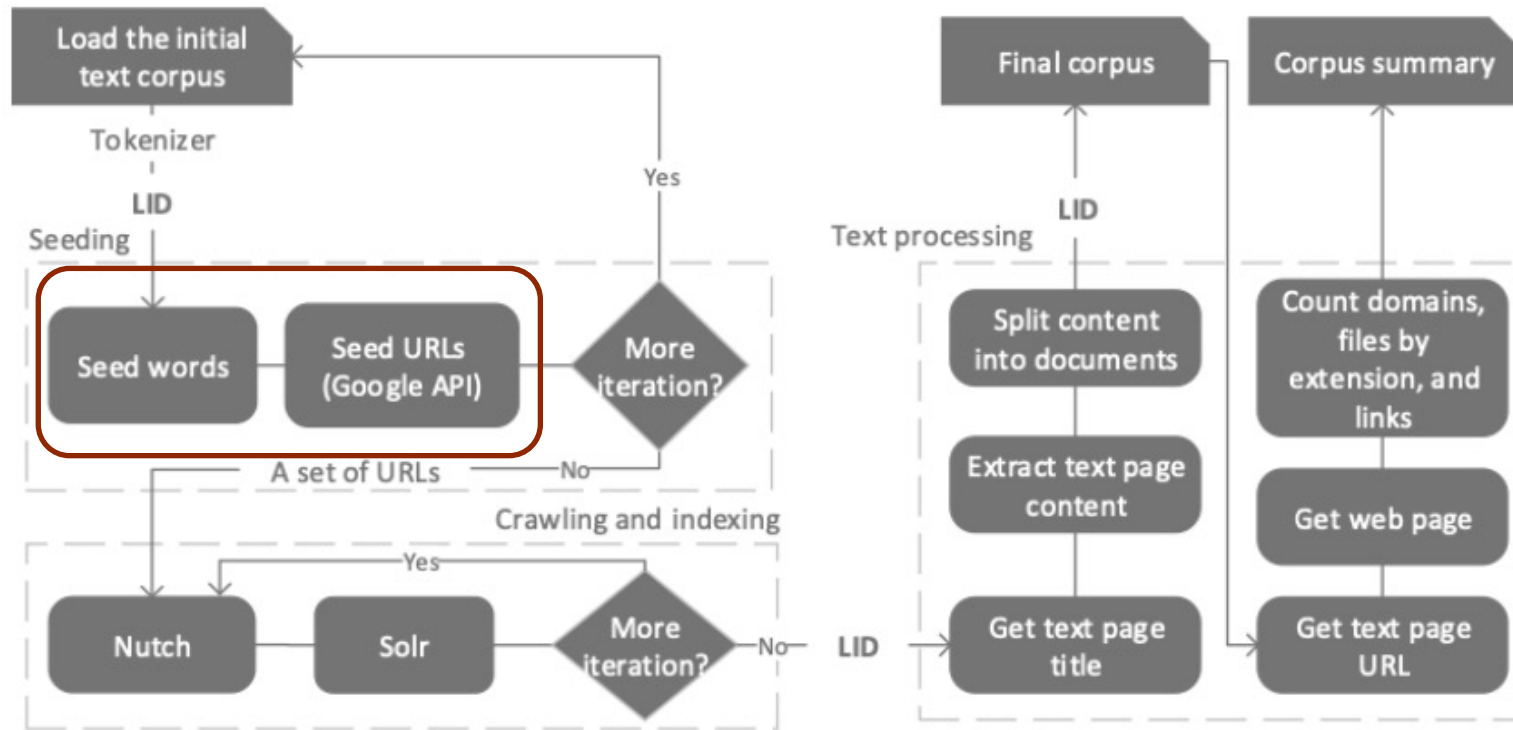
Labadain Crawler

General architecture



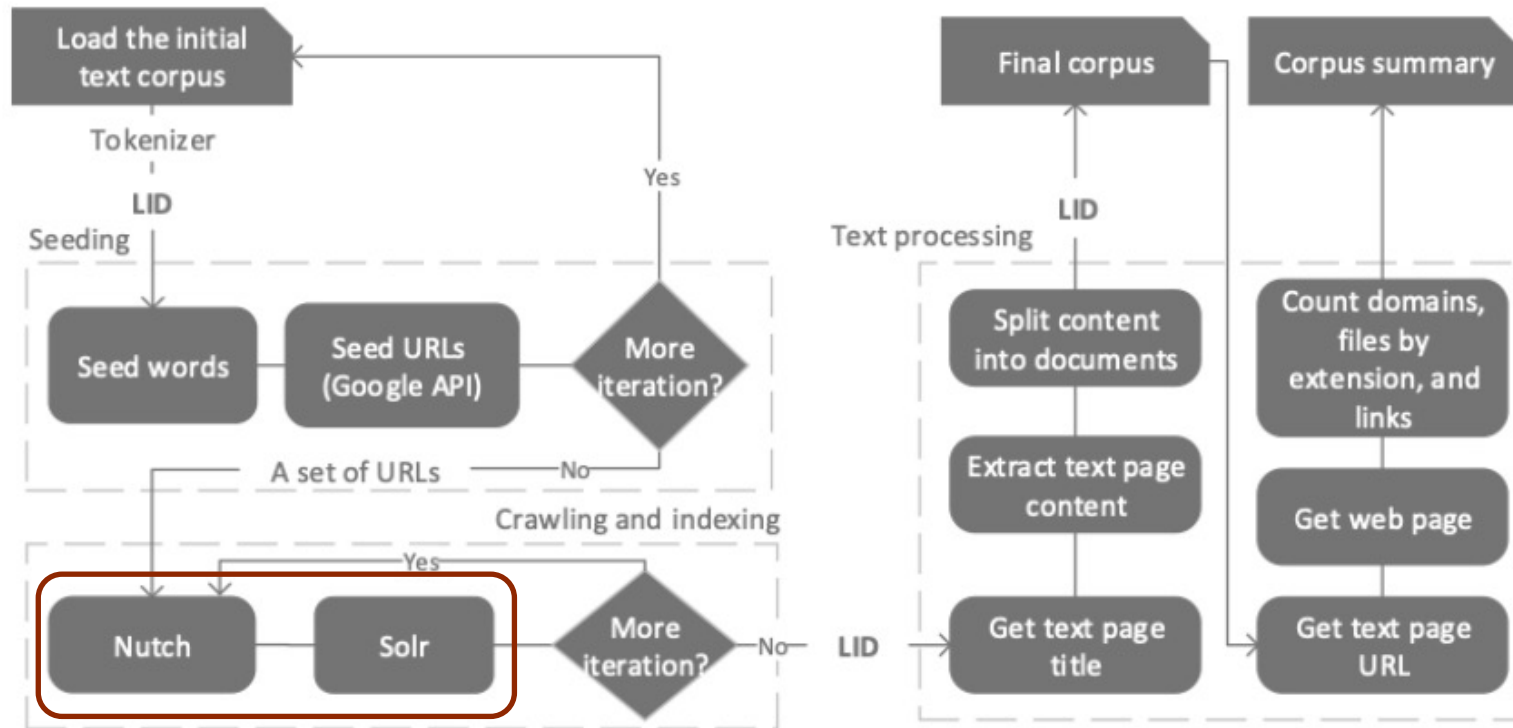
Labadain Crawler

General architecture



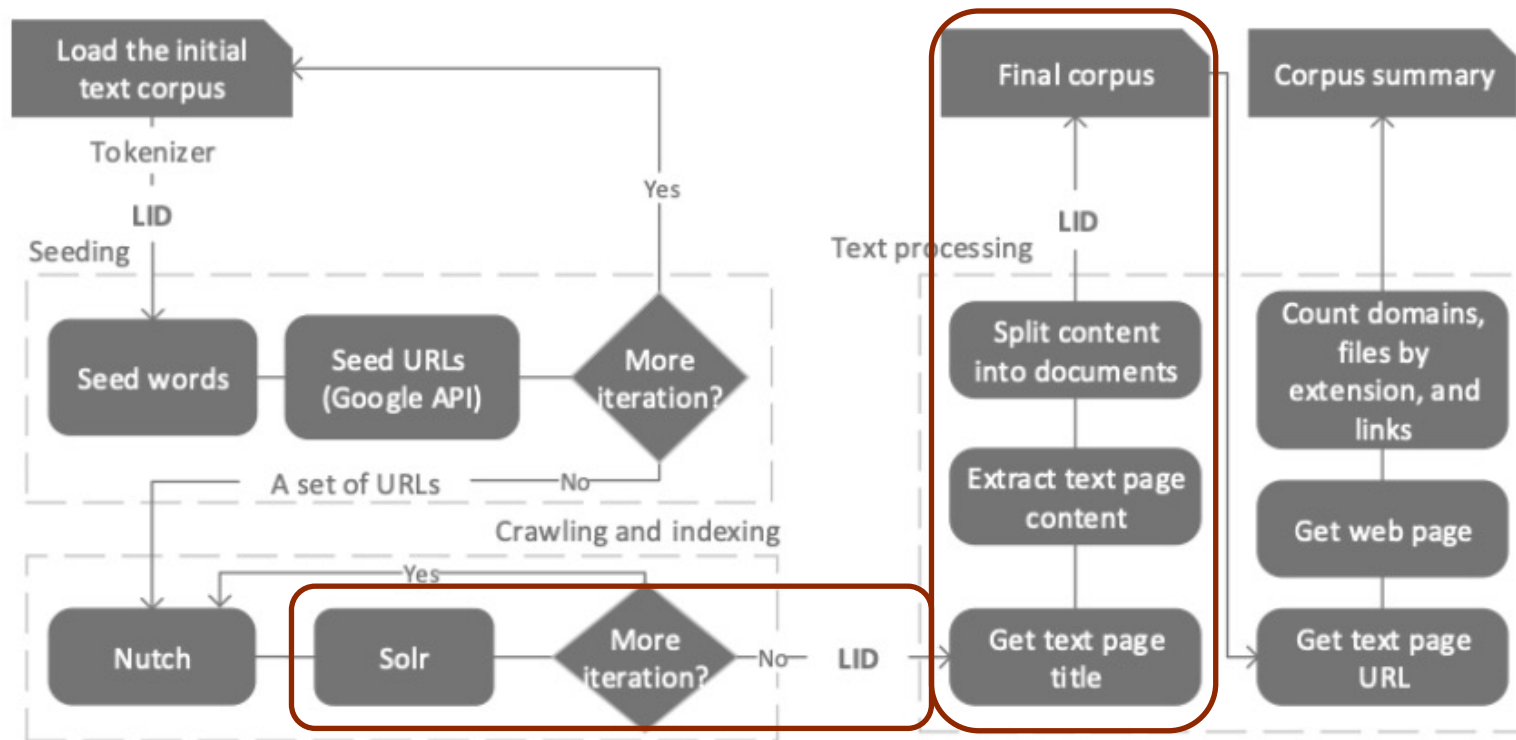
Labadain Crawler

General architecture



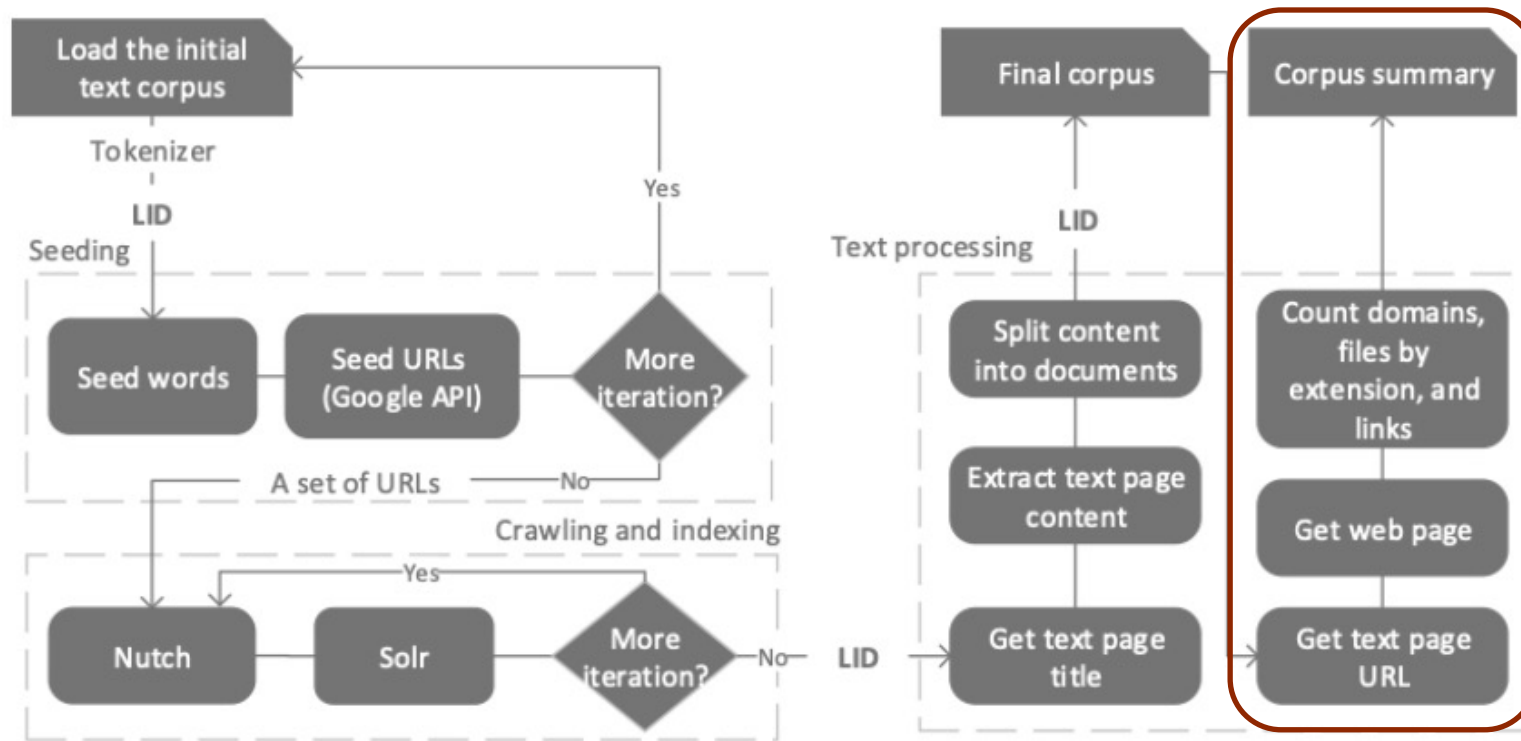
Labadain Crawler

General architecture



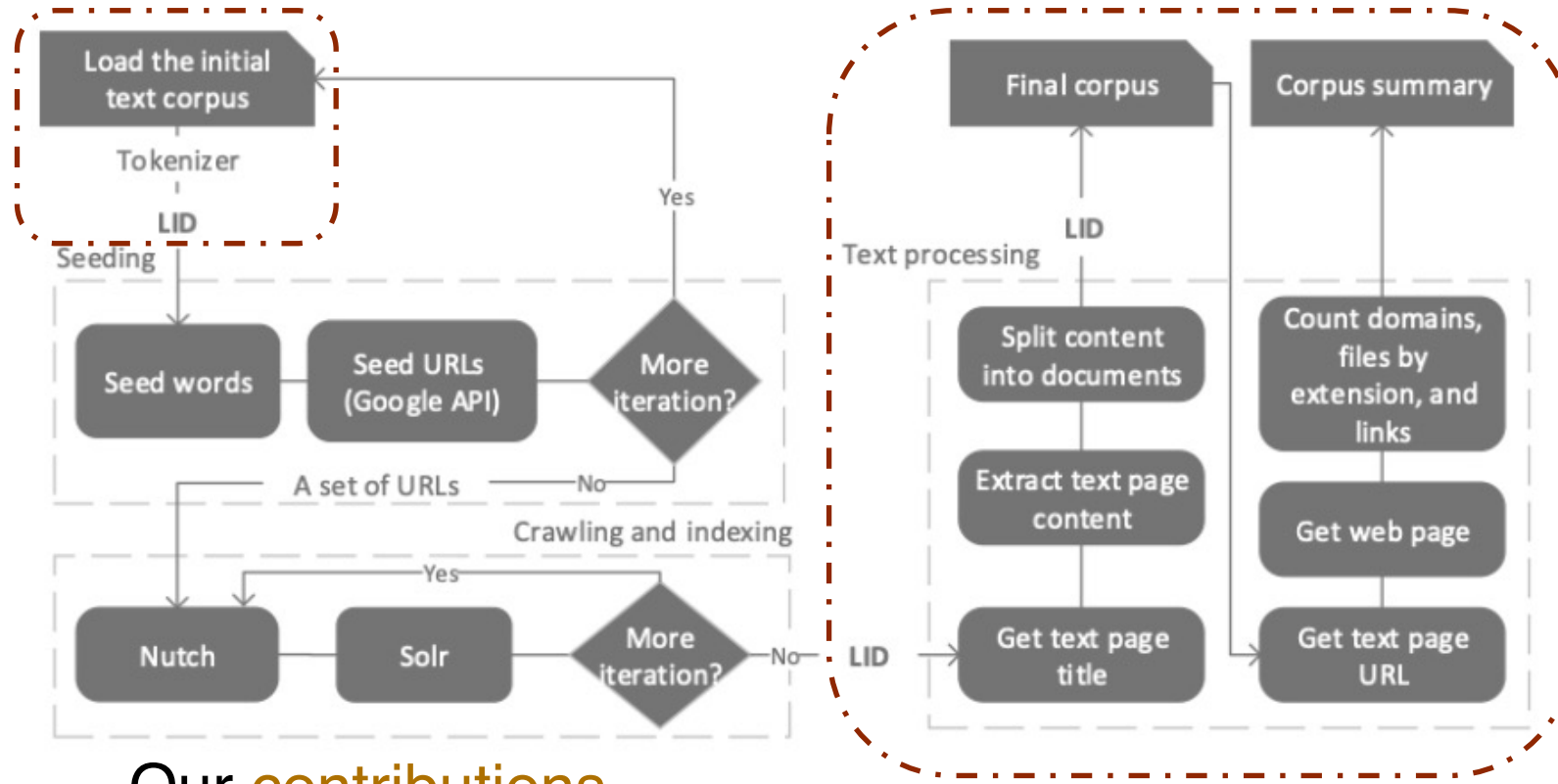
Labadain Crawler

General architecture



Labadain Crawler

General architecture



An Experiment with Tetun

Pipeline configurations:

- Random sampling 10% from an initial corpus of 3500 docs.
- Tetun Word Tokenizer.
- Tetun LID threshold: 0.95.
- Ten seeding repetitions.
- Fifteen crawling repetitions.
- Five crawling depth limit.

Computational specs:

- Linux Ubuntu VM.
- 16GB of RAM.
- 150GB of HDD.
- Single socket CPU with 4 cores.

Experimental Results

After running for approximately 46 hours:

Acquired a text corpus containing:

Corpus content	Total
Text pages	22,392
Sentences	321,721
Tokens*	9,393,499

Including text extracted from:

- 589 PDF files.
- 13 PowerPoint files.

Total of text pages by sources

Data source category	#text pages	Proportion
Online newspapers	16,509	73.73%
Gov. institutions	3,222	14.39%
Non-gov. institutions	1,965	8.78%
Educational institutions	388	1.73%
Blogs and Forums	117	0.52%
Wikipedia	105	0.47%
Personal Pages	57	0.25%
Banks and courts	29	0.13%

Total number of text pages by domains

Domain	#text pages	Proportion
.tl	10,741	47.97%
.com	9,859	44.03%
.org	1,000	4.47%
Others	792	3.54%

Evaluations

Assess the **text page content**:

- Evaluating the text page contents from their **domain names**.
- A total of **22 text pages** (0.01% of the corpus) from **20 domains** did not containing Tetun text.
- The 22 non-Tetun text pages demonstrated that the **LID model was bias** towards “**COVID-19**” and “**Timor-Leste**” terms in short text.

Evaluations

Assess text page quality adopting “quality at a glance” technique:

- Six native Tetun speaker students assessed the text pages’ quality.
- Each student evaluated 50 text pages, a total of 300 documents being assessed.

Evaluations

Factors assessed in the text page quality:

- Quality of the text page titles and contents
- Noise
- Recency and relevancy
- Overall assessment

Evaluations

Assessment results:

Quality metric	Description	#text pages	Proportion
Text page title quality	The text page title is in Tetun	300	100.00%
Text page content quality	The text page contains one or more articles	295	98.33%
Noise	The text page contains clean text	300	100.00%
Recency and Relevancy	Relevant content for present-day usage	278	92.67%
Overall Assessment	Diverse sources with high-quality content	295	98.33%

- Text page titles are in Tetun and contain clean content.
- Over 98% of the text pages contain one or more articles.
- Approx. 93% of the text pages contain information of the latest five years, or older but remain relevant for today's use.
- Overall, the text corpus contains diverse sources and high-quality content.

Conclusions

- The Labadain Crawler has **proven effective** when tested with Tetun.
- The LID model is accurately classified text containing **predominantly Portuguese loanwords** as Tetun.
- The LID model bias can be attributed to the **higher frequency** of these terms **in Tetun dataset** used for training, while being absent in other datasets.
- The Labadain Crawler can be **easily customized** and **adapted** to other LRLs.

Future Work -> Tetun Dataset

Labadain-30k+: A Monolingual Tetun Document-Level Audited Dataset

Gabriel de Jesus, Sérgio Nunes

INESC TEC and Faculty of Engineering of the University of Porto (FEUP)
Rua Dr. Roberto Frias, 4200-465 Porto, Portugal
gabriel.jesus@inesctec.pt, sergio.nunes@fe.up.pt

Abstract

This paper introduces Labadain-30k+, a monolingual dataset comprising 33.6k documents in Tetun, a low-resource language spoken in Timor-Leste. The dataset was acquired through web crawling and augmented with Wikipedia documents released by Wikimedia. Both sets of documents underwent thorough manual audits at the document level by native Tetun speakers, resulting in the construction of a Tetun text dataset well-suited for a variety of natural language processing and information retrieval tasks. This dataset was employed to conduct a comprehensive content analysis aimed at providing a nuanced understanding of document composition and the evolution of Tetun documents on the web. The analysis revealed that news articles constitute the predominant documents within the dataset, accounting for 89.87% of the total, followed by Wikipedia documents at 4.34%, and legal and governmental documents at 3.65%, among others. Notably, there was a substantial increase in the number of documents in 2020, indicating 11.75 percentage points rise in document quantity, compared to an average of 4.76 percentage points per year from 2001 to 2023. Moreover, the year 2017, marked by the increased popularity of online news in Tetun, served as a threshold for analyzing the evolution of document writing on the web pre- and post-2017, specifically regarding vocabulary usage. Surprisingly, this analysis showed a significant increase of 6.12 percentage points in the Tetun written adhering to the Tetun official standard. Additionally, the persistence of Portuguese loanwords in that trajectory remained evident, reflecting an increase of 5.09 percentage points.

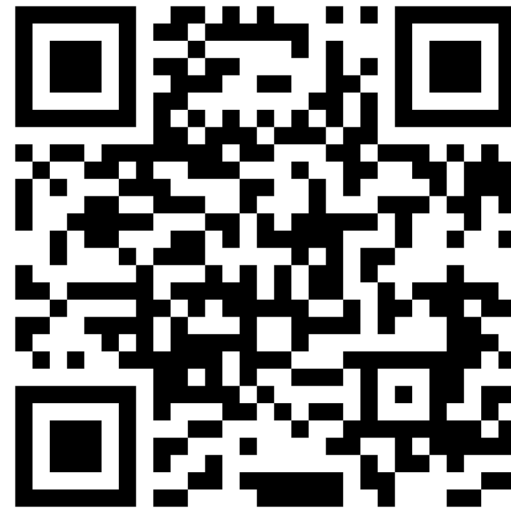
Keywords: Low-resource language, Tetun, Text dataset, Corpus content analysis.

Scan me to access the dataset!



Thank You

Scan here to access the
Labadain Crawler code!



Data Collection Pipeline for Low-Resource Languages: A Case Study on Constructing a **Tetun** Text Corpus

Gabriel de Jesus, Sérgio Nunes

INESC TEC / Faculty of Engineering, University of Porto (FEUP)

The 2024 Joint International Conference on Computational Linguistics,
Language Resources, and Evaluation

Lingotto Conference Centre, Turin, Italy
20-15 May, 2024