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 Lisguistics, Language Resources, and Evaluation

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

U.PORTO
FEUP FACULDADE DE ENGENHARIA UNIVERSIDADE DO PORTO

fct Fundação para a Ciência e a Tecnologia

INESCTEC TECHNOLOGY & SCIENCE | ASSOCIATE LABORATORY PORTUGAL

Outline

- Context and Motivation
- Related Work
- Propose Solutions
- Tetun Tokenizer
- Tetun LID
- Labadain Crawler
- An Experiment with Tetun
- Results and Evaluations
- Discussions
- Conclusions & Future Work

Context and Motivation

- Text corpora are crucial for the development 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.

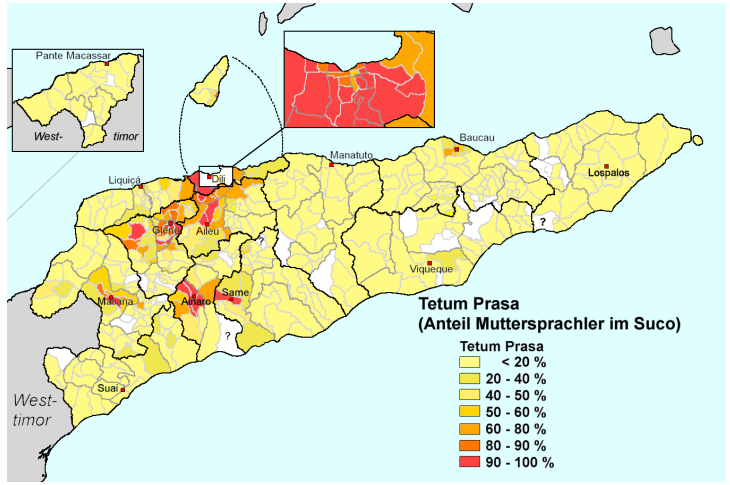
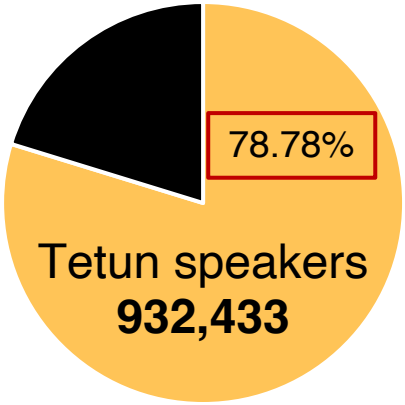


Image source: Wikipedia

- One of Timor-Leste's Official languages alongside Portuguese



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.

Mikel Artetxe, Itziar Aldabe, Rodrigo Agerri, Olatz Perez-de-Viñaspre, and Aitor Soroa. 2022. Does corpus quality really matter for low- resource languages? In Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022, pages 7383–7390. Association for Computational Linguistics.

Related Work

Korner et al. (2022):

Crawling and Collaborative:

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

Implementation:

Constructing corpora for 258 LRLs.

Challenges:

Motivating community to participate.

Erik Körner, Felix Helfer, Christopher Schröder, Thomas Eckart, and Dirk Goldhahn. 2022. Crawling under-resourced languages – a portal for community-contributed corpus collection. In Proceedings of the 1st Workshop on Dataset Creation for Lower-Resourced Languages (DCLRL) @LREC2022, Marseille, 24 June 2022. European Language Resources Association (ELRA).

Related Work

Linder et al. (2020):

SwissCrawl:

A web crawling tool developed from scratch.

Implementation:

Constructing a text corpus for Swiss German.

Challenges:

Limited availability of documentations.

Lucy Linder, Michael Jungo, Jean Hennebert, Claudiu Cristian Musat, and Andreas Fischer. 2020. Automatic creation of text corpora for low-resource languages from the internet: The case of swiss german. In Proceedings of The 12th Language Resources and Evaluation Conference, LREC 2020, Marseille, France, May 11-16, 2020, pages 2706–2711. European Language Resources Association.

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.

Guillaume Wenzek, Marie-Anne Lachaux, Alexis Conneau, Vishrav Chaudhary, Francisco Guzmán, Armand Joulin, and Edouard Grave. 2020a. Ccnet: Extracting high quality monolingual datasets from web crawl data. In LREC, pages 4003–4012. European Language Resources Association.

Propose Solutions

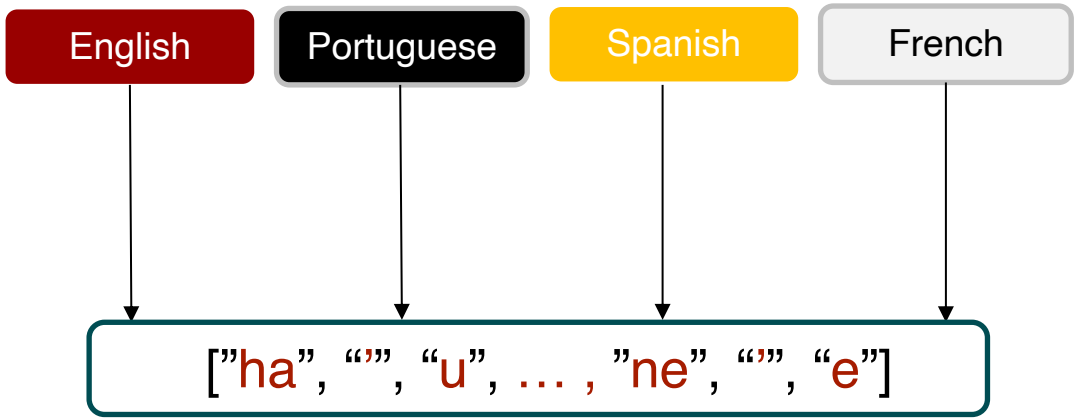
Labadain Crawler - A data collection pipeline that relies on three key components:

- 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
```

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.

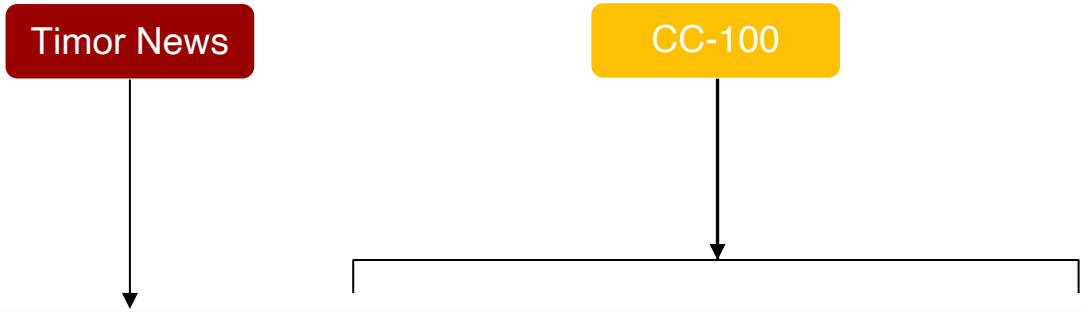
Tetun Tokenizer

Evaluations:

- Five Timorese volunteer students evaluated each tokenization techniques.
- Each evaluated a minimum of three text samples (200 to 250 words) that were collected from the web.
- Evaluators reported that all 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

https://www.timornews.tl
https://data.statmt.org/cc-100/

Tetun Language Identification

Training and Evaluation

Input: sentences

Labels: languages

Training set: 70% Dev. set: 15%

Test set: 15%

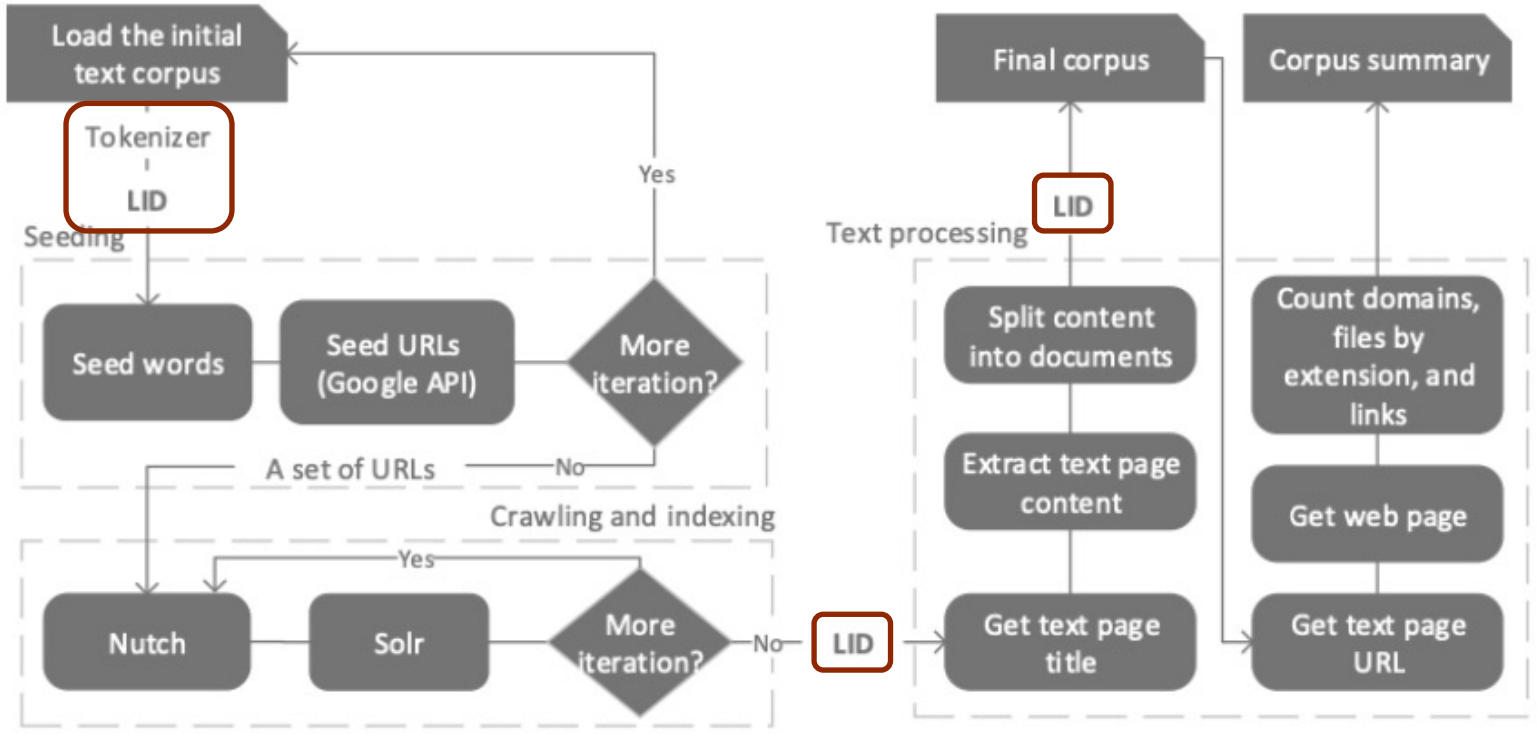
Model	Character n-gram						Word n-gram		
	1	2	3	4	5	6	1	2	3
SVM	0.9822	0.9954	0.9974	0.9975	0.9974	0.9968	0.9953	0.9689	0.8397
LR	0.9812	0.9953	0.9970	0.9975	0.9971	0.9960	0.9930	0.9522	0.8107
MNB	0.9452	0.9918	0.9967	0.9977	**0.9981**	0.9979	0.9973	0.9755	0.7806

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

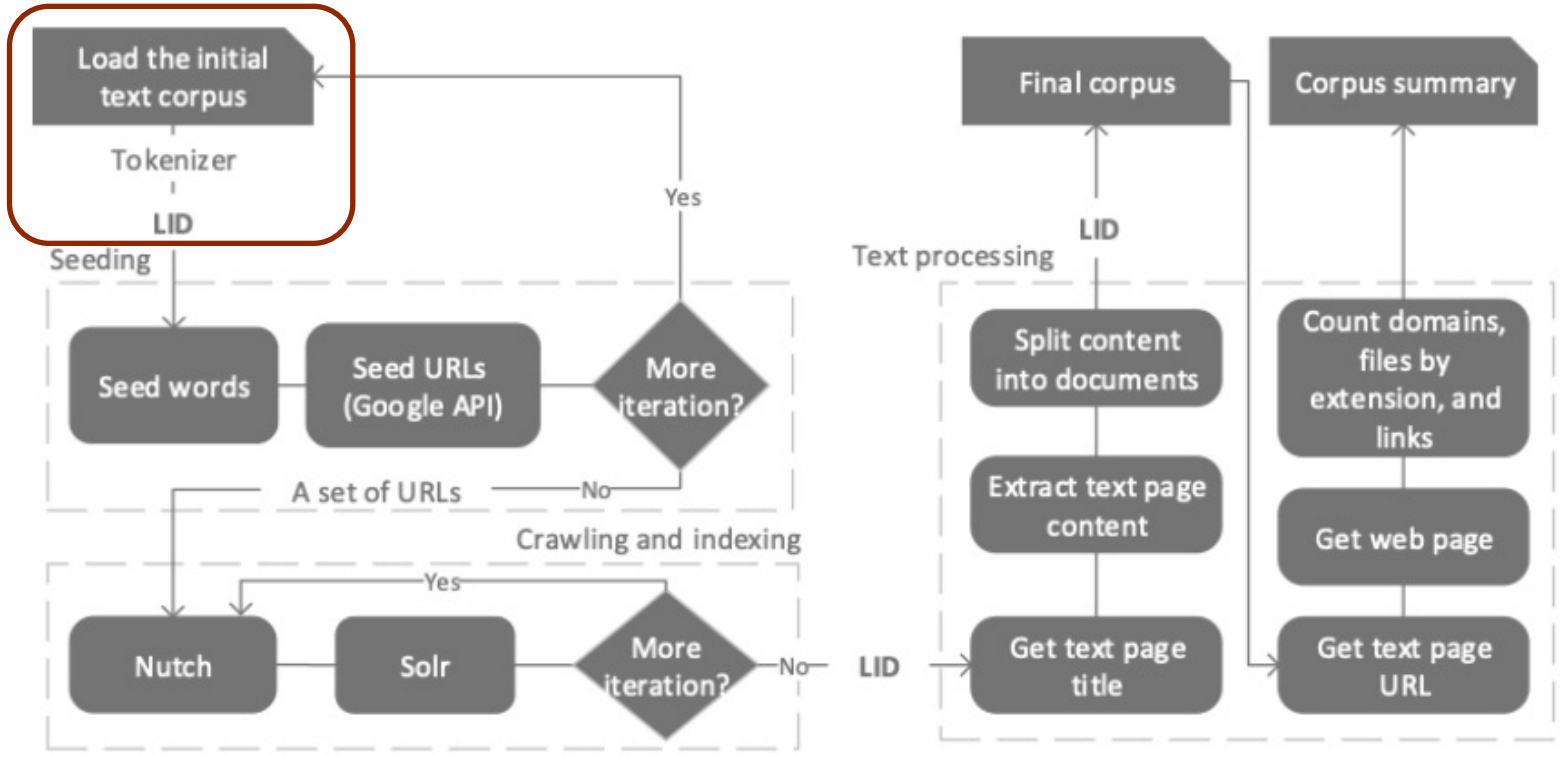
Labadain Crawler

General architecture



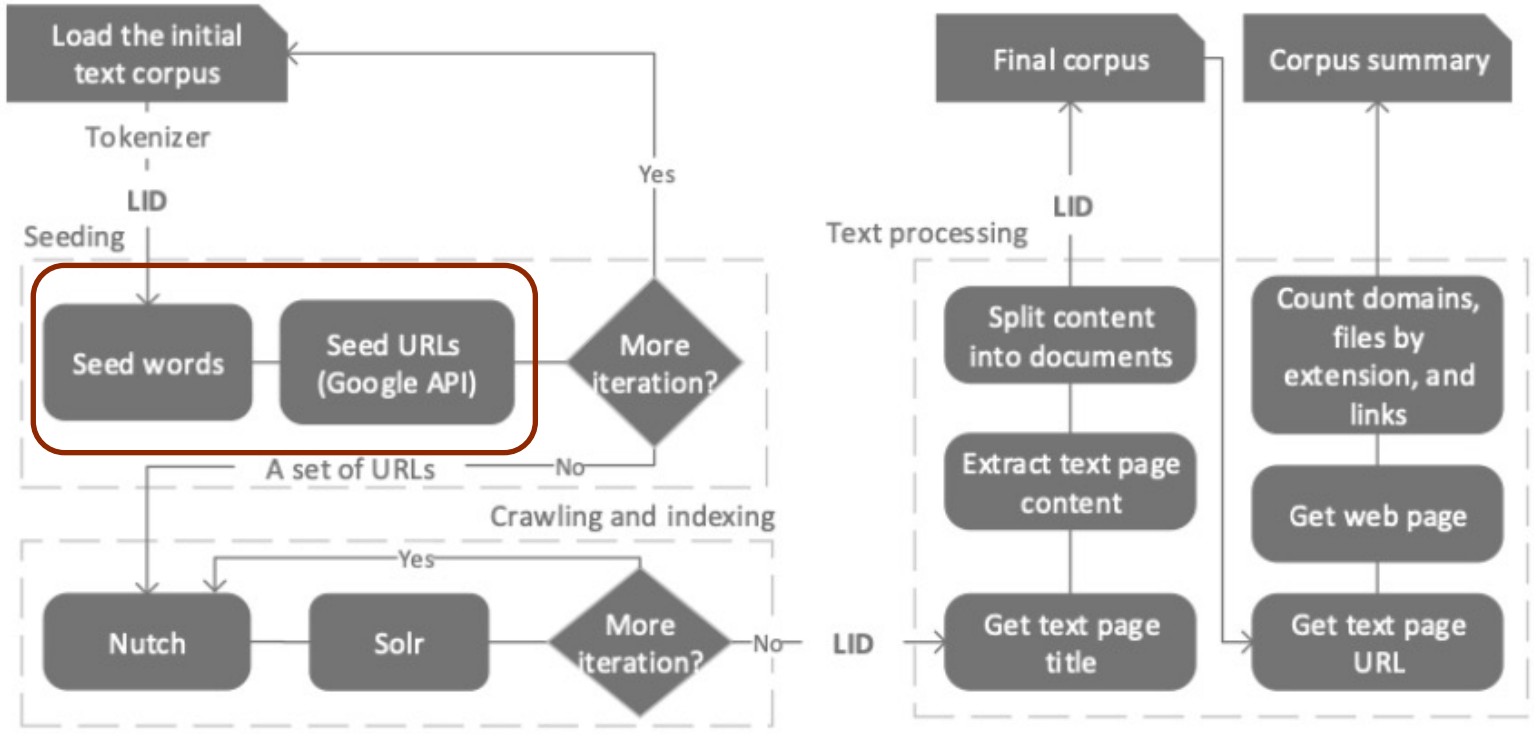
Labadain Crawler

General architecture



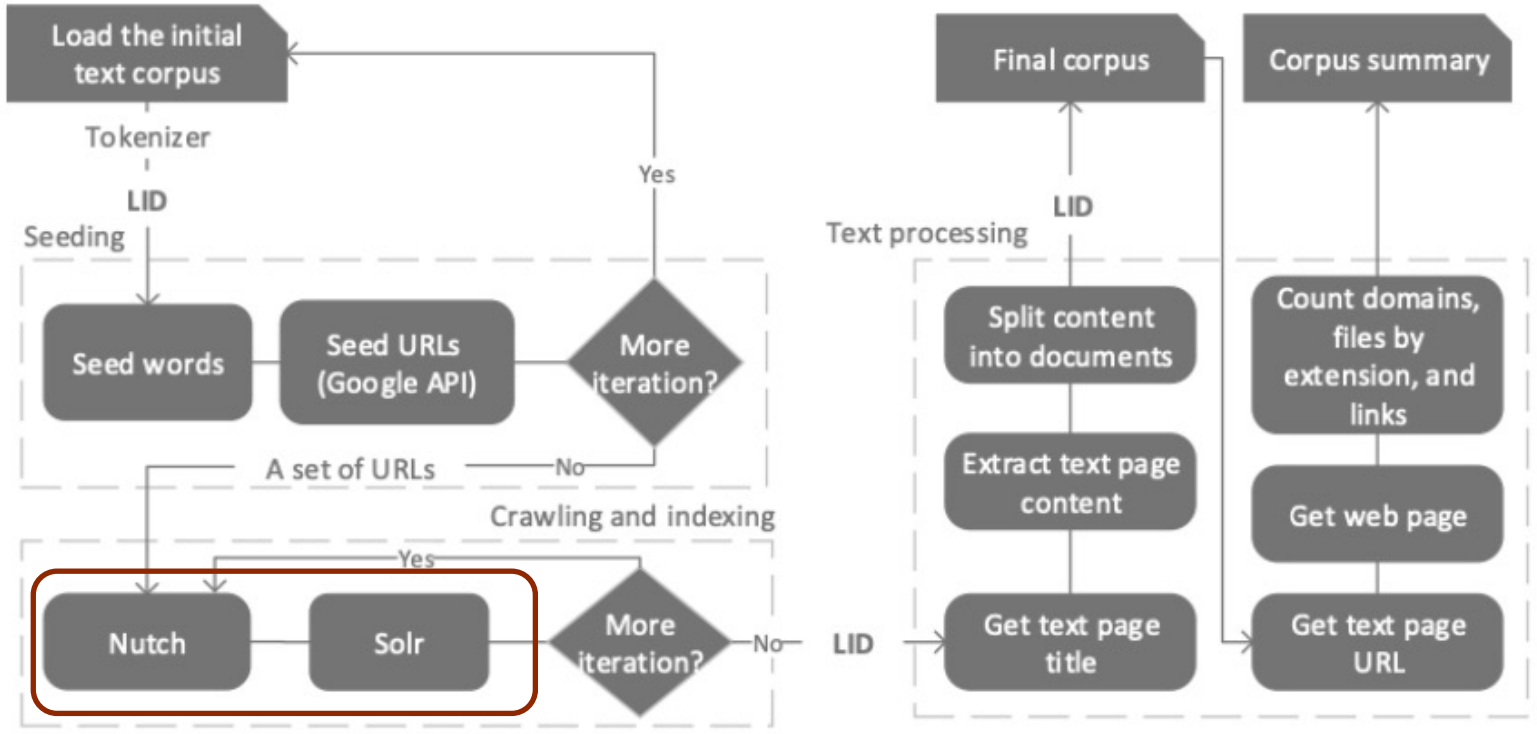
Labadain Crawler

General architecture



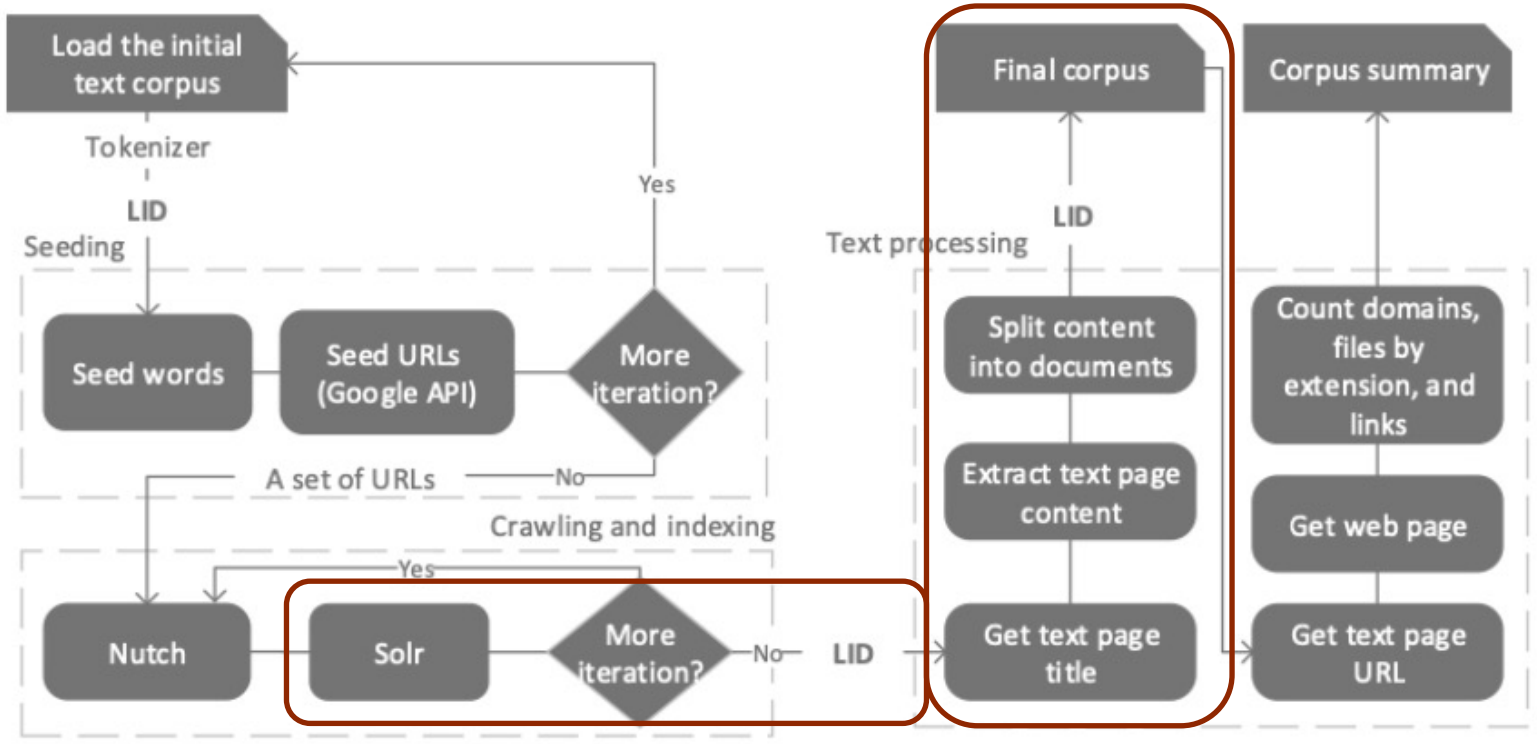
Labadain Crawler

General architecture



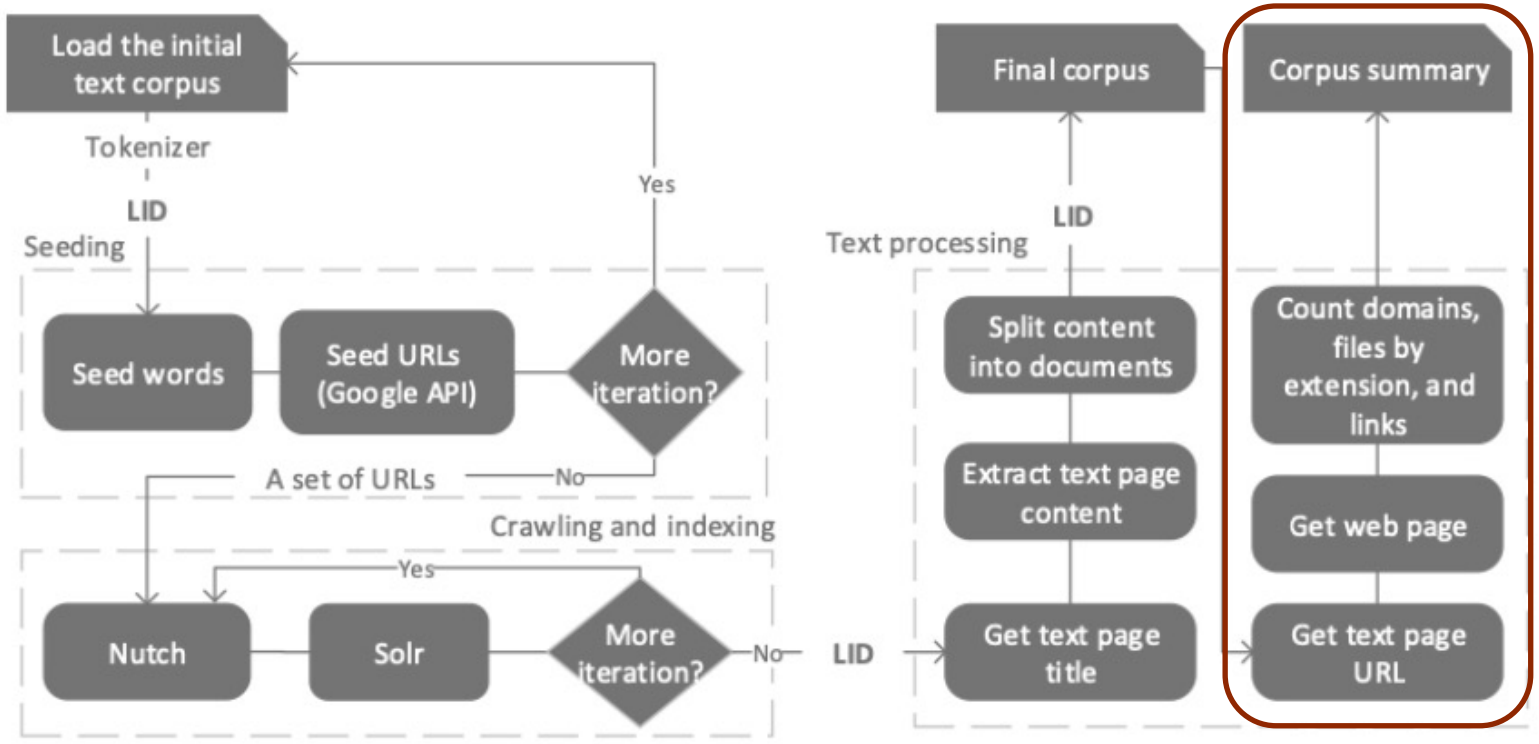
Labadain Crawler

General architecture



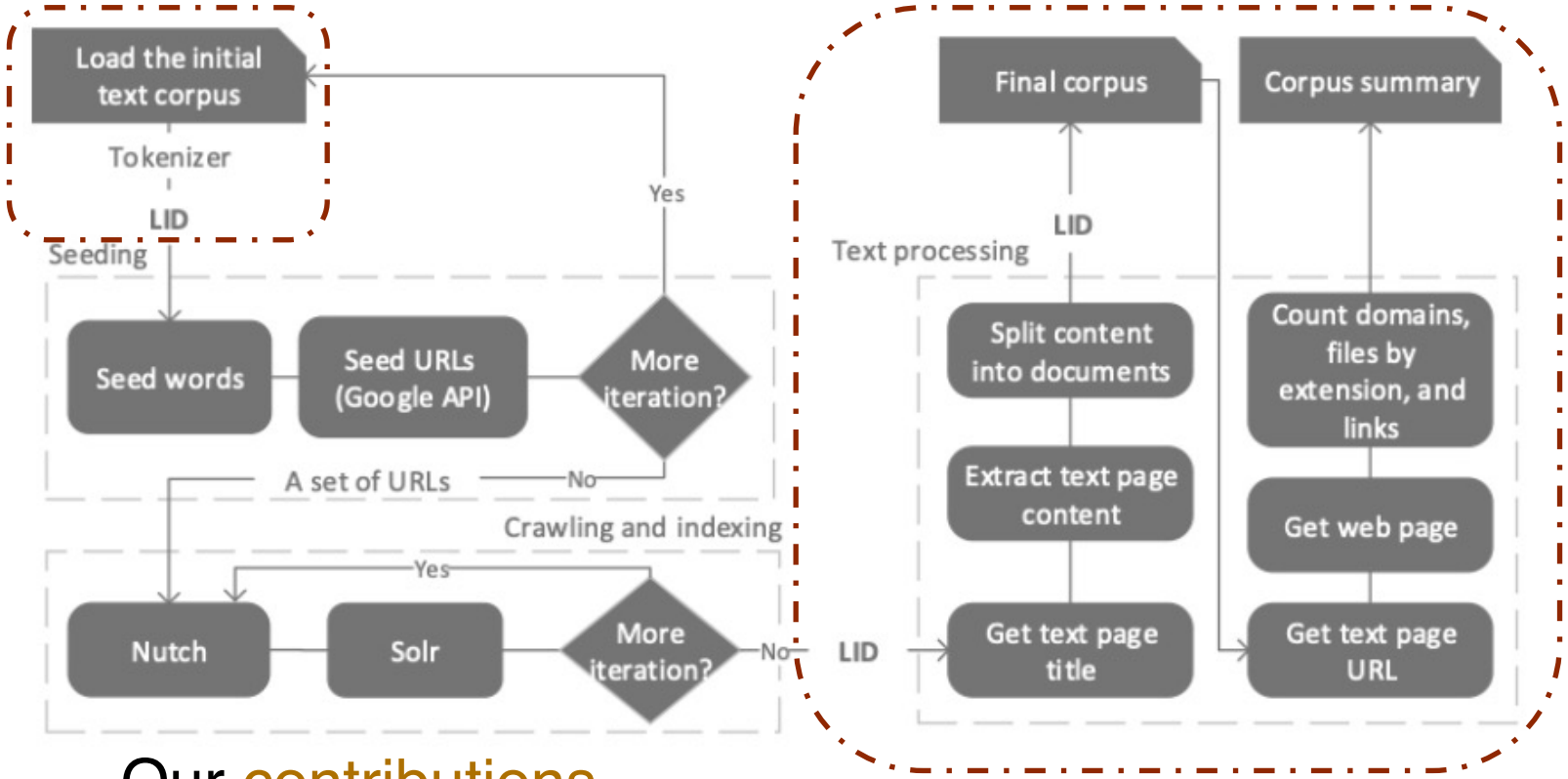
Labadain Crawler

General architecture



Labadain Crawler

General architecture



--- Our contributions

An Experiment with Tetun

Pipeline configurations:

- An initial text of 3500 docs (random sampling 10%).
- Tetun Word Tokenizer.
- Tetun LID threshold score: 0.95.
- Ten seeding repetitions.
- Fifteen crawling repetitions.

Computational specs:

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

Experimental Results

After running for ~46 hours:

Acquired a text corpus comprising:

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

General assessments:

- Evaluating text pages quality from the domain names.
- A LID model biased towards 22 text pages, corresponding to 0.01% of the corpus.

Text pages quality: adopting “quality at a glance” approach

- Five students conducted the assessments.
- Evaluating a total of 300 text pages randomly sampling.

Kreutzer et al.. 2022. Quality at a glance: An audit of web-crawled multilingual datasets. Trans. Assoc. Computational Linguistics, 10:50–72.

Evaluations

Adopting “quality at a glance” approach

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%

Kreutzer et al.. 2022. Quality at a glance: An audit of web-crawled multilingual datasets. Trans. Assoc. Computational Linguistics, 10:50–72.

Evaluations

Adopting “quality at a glance” approach

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%

Kreutzer et al.. 2022. Quality at a glance: An audit of web-crawled multilingual datasets. Trans. Assoc. Computational Linguistics, 10:50–72.

Evaluations

Adopting “quality at a glance” approach

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%

Kreutzer et al.. 2022. Quality at a glance: An audit of web-crawled multilingual datasets. Trans. Assoc. Computational Linguistics, 10:50–72.

Evaluations

Adopting “quality at a glance” approach

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%

Kreutzer et al.. 2022. Quality at a glance: An audit of web-crawled multilingual datasets. Trans. Assoc. Computational Linguistics, 10:50–72.

Evaluations

Adopting “quality at a glance” approach

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%

Kreutzer et al.. 2022. Quality at a glance: An audit of web-crawled multilingual datasets. Trans. Assoc. Computational Linguistics, 10:50–72.

Discussions

- The Tetun tokenizer can be adapted to other languages, e.g., ENG, PT and ID.
- The LID model is accurately classified Portuguese loanwords as Tetun.
- The LID model is biased towards terms of COVID-19 and Timor-Leste in short texts.

Conclusions and Future Work

- The Labadain Crawler can be operated even with limited computational resources.
- The Labadain Crawler's effectiveness depends on robust language processing components.
- The Labadain Crawler can be easily customized and adapted to other LRLs.
- In future work, we will mitigate the LID model bias issues to improve the performance.

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 Lisguistics, Language Resources, and Evaluation

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

U.PORTO
FEUP FACULDADE DE ENGENHARIA UNIVERSIDADE DO PORTO

fct Fundação para a Ciência e a Tecnologia

INESCTEC TECHNOLOGY & SCIENCE | ASSOCIATE LABORATORY PORTUGAL