Sri Lanka Document Datasets: A Large-Scale, Multilingual Resource for Law, News, and Policy

Nuwan I. Senaratna Independent Researcher
nuwans@alumni.stanford.edu

October 4, 2025

Abstract

We present a collection of open, machine-readable document datasets covering parliamentary proceedings, legal judgments, government publications, news, and tourism statistics from Sri Lanka. The collection currently comprises ∼215,552 documents (59.6 GB) across 13 datasets in Sinhala, Tamil, and English, updated daily and mirrored on GitHub and Hugging Face. These datasets aim to support research in computational linguistics, legal analytics, socio-political studies, and multilingual natural language processing. We describe the sources, collection pipeline, formats, and potential use cases, while discussing licensing and ethical considerations.