Next generation AI-based ecosystem for Noolaham Foundation - Second Phase

Project Title	Large Scale Annotation, Storage and Analysis of Digitised Sri Lankan Tamil Content
Project Location	Sri Lanka
Implementing Institution	Noolaham Foundation
Total Budget	$31428
Duration	-
Team	Shaseevan Ganeshananthan, Prashanth Srinivasan, Saatviga Sudhahar

Project Summary

The project aims to do large scale annotation, storage and analysis of Sri Lankan Tamil content. This is related to the field of semantic culturomics in which researcher's data mine large digital archives to investigate cultural phenomena reflected in language and word usage. It is a form of computational lexicology that studies human behavior and cultural trends through the quantitative analysis of digitised texts. The underlying data is from Noolaham Foundation, a Digital Archive and a Digital Library undertaking the critical work of documenting, digitally preserving and providing free and open access to knowledge bases and cultural heritage of Sri Lankan Tamil speaking communities. The archive contains digitised text from Sri Lankan newspapers, books, magazines, pamphlets etc. from various sources totaling up to approximately 100,000+ documents. It also includes a web archive and born-digital data in text format which would be included in our pipeline.

Goals

The goal of the project is to build an ecosystem that analyses large scale Sri Lankan Tamil content using natural language processing methodologies and generative AI.

Objectives

- Converting digitized content from Noolaham project to text with meta-data tagging. This includes converting newspapers, books, and magazines, pamphlets from pdfs and images to text. We do Layout analysis of newspapers, Optical character recognition, Building document type storage, XML conversion of text,
- Auto-labelling of documents starting from manual annotation and training a model to automatically annotate content
- User Interface to display extracted text with annotations for all categories in Noolaham's content
- Building language processing resources using natural language processing resources and building Knowledge engineering capabilities.
- Building a generative AI based tool to query Noolaham's content, similar to ChatGPT.

Conclusion

As part of the language preprocessing layer of the ecosystem we have completed the development of the digital content conversion pipeline. As next steps layout analysis, ocr and xml conversion has to be run on the whole of Noolaham's content to convert it to raw text. This requires significant resources and budget related to server cost and running time. I will submit a separate proposal specifying those requirements.

Separate efforts have to be initiated to build the processing resources layer and the knowledge engineering layer.

Budget

Noolaham Total Digitized material - 150,926

Total Pages - 5,509,494

The cost breakdown for all phases are given below.

Details of Expenditure	Unit cost	Unit	Quantity	Total
Language Pre-processing				
Running the Digital Content Conversion Pipeline				
Layout Analysis Optical Character Recognition XML Conversion	Aws ec2 - i4g.4xlarge on demand pricing	$1.24 per hr for 3444 hrs	4 instances	$17080
Input data storage	Aws s3 50TB/month 208TB needed	$0.023 per GB	4 months	$4784
Converted data storage in TXT	Aws ec2 - i4g.4xlarge Storage pricing	50GB	12 months	$50
Auto-labelling content				
Metadata creation - label content	Annotators	60000 LKR per person per month	3 persons for 5 months	$3006
Auto-topic tagging model development	ML Researcher	100000 LKR per person per month	2 months	$668
UI for Displaying TXT content				
User Interface Development	Software Engineer	100000 LKR per person per month	2 months	$668
Processing Resource Layer				

Gap Analysis & development of tools	NLP Researcher (Intern)	50000 LKR per person	6 months	$1002
Knowledge Engineering				
Building a TamilGPT	NLP Researcher	150000 LKR	5 months	$2500
- Infrastructure cost for Maintenance of TamilGPT	AWS ec2 G4dn.8xlarge or Self-built server	$1.306 per hr	12 months	$11440 or $6000
Knowledge engineering data analysis	NLP Researcher	100000 LKR per person	5 months	$1670
Sub Total				
				$31428
Grand Total				**$31428**