Title: From Uncertainty to Action: Recalibrating Digital Humanities Methods and Tools for Non-standard British Colonial South India Data

Principal Investigator (PI):

Shanmugapriya T
Assistant Professor
IIT(ISM) Dhanbad

bhavani	NN	dharapuram	NN	cavary	JJ	coimabatc	NN
bhavdni	NN	dharipuram	JJ	cavery	NN	coimbator	NN
bhavgni	NN	dharmapui	NN	kavari	NN	coimbetor	RB
bhavini	NN	dharmapuri	NN	kavary	JJ		
bhavitni	NN	dharmavaram	NN	kaveri	NN		
bhavttni	NN	dharwar	JJ	kavery	NN		
bhawini	NN	dhdripuram	NN				
bhaysni	NN	dhebri	NN				
bhivani	JJ	dhtirapuram	NN				

Sample inconsistent spellings of toponyms extracted from Frederick Nicholson's *Manual Of The Coimbatore District in the Presidency of Madras* (1887)

Overview of the project:

The research aims to recalibrate digital humanities methods and tools for effectively analyzing uncertain and complex colonial historical data to extract fuzzy toponyms. Additionally, I investigate how advanced digital techniques can contribute to the development of a novel framework rooted in historical data from the global South. To address these inquiries, I employ a case study approach, with a particular focus on the British South India corpus, utilizing advanced nature language processing techniques such as BERT Named Entities Recognition (NER) and DeezyMatch. The primary goal of this research is to extract toponyms from the selected British India colonial corpus (1799-1947).

Research Methodology:

The proposed method consists of two stages. In the first stage, the focus is on identifying location entities by using BERT Named Entity Recognition (NER) (Devlin et al. 2018). This NER aids to identify concealed toponyms by learning from the context. The second stage is dedicated to extracting the fuzzy toponyms. To achieve this, I will utilize DeezyMatch1, a free library developed by Kasra Hosseini, Federico Nanni, and Mariona Coll Ardanuy (Hosseini, Nanni, and Ardanuy 2020). DeezyMatch is specifically designed for fuzzy string matching and toponym extraction, providing candidate ranking. To generate the training dataset for string pairs, I will compile alternate names of places in South India besides the trained data set from BER. By learning similar transformations as those in the training set, DeezyMatch should be capable of applying this learning to unseen variations of toponyms.

Deliverables:

First Phase: Data Collation: During this phase, the primary focus will be data collation and cleaning. As the existing data is insufficient for machine learning and limited to a specific region in South India, I will gather digitised documents from open-access online archives, such as the Internet Archive, Google Books, the National Library of Calcutta, and other online libraries. Pre-processing techniques will be applied to convert the collected texts into a machine-readable format.

Expected Deliverable: A comprehensive machine-readable dataset of British South India.

Second Phase: The main objective is to identify and extract toponyms using advanced programming techniques such as BERT NER and DeezyMatch, as outlined in the methodology.

Expected Deliverables: The list of toponyms in British South India, two journal papers, and the dissemination of results, experimentation, and lessons learned from the experiments through participation in Digital Humanities conferences, workshops, seminars, and symposiums, both within and outside Indian at the Institute of Technology Dhanbad.