CS 613: NLP

Assignment 1: Data Preparation and De-duplication

1. Introduction

In this project, we aimed to curate, process, and clean a dataset of language **MALAYALAM** for language processing, focusing on ensuring data quality and suitability for subsequent analysis. The curated datasets were gathered from multiple publicly accessible sources, cleaned of inappropriate content, and deduplicated to ensure the integrity and usability of the dataset.

2. Dataset Curation

The datasets were sourced from various online platforms, including Wikipedia, other freely accessible websites and existing datasets available on kaggle, hugging face,etc. The goal was to gather a diverse set of textual data for MALAYALAM.

3. Data Cleaning

The collected dataset underwent a cleaning process to ensure the removal of inappropriate content, such as articles containing bad words, pornographic material, hate speech, and abusive language. The cleaning was conducted using custom-developed tools and existing dictionaries for bad words (the list is present in codebase present on github). Here is a table that provides an overview of the dataset sources, their respective volumes in GBs, and the total number of articles collected from each source before and after cleaning:

4. Deduplication

After cleaning, the next step was to remove duplicate articles to improve the dataset's quality. This was achieved using both the provided codebase and additional robust deduplication techniques.

All techniques present in-
https://colab.research.google.com/drive/1tN0Pf66k65-1uyx3P1g6sbyeYJVhrsBo?usp=sharing

Note - Due to connectivity issues in the server and space restrictions on local machines we implemented deduplication on a selected dataset which we curated keeping in mind the source and chances of two sources having duplicate files.

6. Conclusion

This project involved the comprehensive curation, cleaning, and deduplication of a language dataset. By gathering data from various sources and applying rigorous cleaning and deduplication processes, we ensured the dataset's quality for further use in language modeling.

The results show a significant reduction in the dataset's volume while maintaining a high level of content relevance and integrity.

INDIVIDUAL CONTRIBUTIONS:

Team Member Name	Contribution
Heer	Downloaded datasets, compiled the list of bad-word dictionaries, developed and executed the cleaning process to remove articles with inappropriate content, and documented the tools used.
Jiya	Scraped various websites, ensured all articles were saved in `.txt` format, and compiled the table of dataset sources, volumes, and total articles.
Lavanya	Scraped various websites, downloaded datasets and participated in the cleaning process to remove articles with any inappropriate content.
Shrishti	Scraped various websites, removed duplicate articles, implemented robust deduplication techniques, and compiled statistics before and after deduplication.
Utkarsh	Downloaded datasets, removed duplicate articles, implemented robust deduplication techniques, and compiled statistics before and after deduplication.

All the codes are present on the github link.