**REQUIREMENTS**

1. Project Overview

The availability of digital information has made it possible to use digital data which is also the cause of the misuse of the available data Some of the issues associated with the misuse of digital data are Plagiarism detection, ownership identification etc. Plagiarism detection is of particular interest to people in the academia and the publishing sector. Plagiarism means copying thought and text of another person and presenting them as ones's own work. English copy detection systems have been studied since 1990s and some of them can be freely downloaded from the Internet while a copy detection system for Malayalam is not available.

This project is intended to develop the same using the language processing methods like stemming, fuzzy search etc.

1. Project Objectives

1. Improvise the existing stemmer for Malayalam Language

2. Perform Plagiarism detection

3.  Calculate percentage of similarity between two documents.

1. Project Outcome

1. Plagiarism detection in Malayalam Texts

2. Stemmer for Malayalam Language that can be used for research purposes by Language Scientists and Experts

1. Week Plan(10 weeks)

//1 week = 4 hours.

|  |  |  |  |
| --- | --- | --- | --- |
| **Week No** | **Objective** | **Learning Outcome** | **Reference** |
| **1** | Presentation of  Feasibility Study, Requirement Specification, System Analysis | Whether project is economically and technologically  feasible. Minimum hardware and software requirements are analysed. | A Copy detection Method for Malayalam Text Documents  using N-grams Model, Sumam Mary Idicula et al. |
| **2** | System Design | Basic Algorithm for stemming | A Literature Review: Stemming Algorithms for  Indian Languages  M.Thangarasu , Dr.R.Manavalan |
| **3** | System Design | Integrating Jaccard’s similarity function over root words. | Stemmers for Tamil Language:  Performance Analysis  M.Thangarasu |
| **4** | System Coding | Implementing Stemming in Python | SILPA |
| **5** | System Coding | Implementing Stemming in Python | SILPA |
| **6** | System Coding | Implementing Stemming in Python | SILPA |
| **7** | System Coding | Implementing Duplicate Detection | Near Duplicate Document Detection Survey [Bassma S. Alsulami, Maysoon F. Abulkhair, Fathy E. Eassa] |
| **8** | Test Suite Preparation, System Testing, Maintenance | Sample textual inputs are generated and used for testing | Malayalam Articles and Text Excerpts |
| **9** | Documentation | Brief description of functions and algorithms are included in the code. |  |
| **10** | Documentation | Brief description of functions and algorithms are included in the code. |  |