Document Similarity Detection for Indic Languages

Objective:

To implement shingling algorithm of Document Similarity Detection for Indic Languages (Malayalam in particular). Due to the property of inflection and agglutination, Indic languages cannot be treated directly with word to word comparisons. So, a fuzzy search, indic soundex and edit distance technologies should be used. Edit distance technology is used for words which are written similarly, indic soundex technology is used for words which sound similar.

Existing System :

No system is in existence for Document Similarity for Malayalam Language. SILPA framework handles inflections partially, and we intend to improvise it.

Proposed System:

The texts are tokenized and each pair of corresponding token is treated with an approximate comparison algorithm which will take care of inflections (if any). Agglutionations are not in the scope of this project. The results of approximate comparison is used in Shingling Algorithm and the Jaccard's Similarity Formula to calculate the similarity. The words are arranged in ascending order and then tokenized to minimalize the problem of "word position similarity".

Stakeholders : Language Scientists, Media personnels, Literature Experts and students.

Coding Backend : Python and SILPA (Swathanthra Indic Language Processing Applications)

GUI Frontend : PyGTK/PyQt