**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

**Relevance of Project**

Holocene epoch has detected the enormous emergence of Internet document in the World Wide Web.Internet is now a vital factor for day today life in gathering information and extracting useful information from web pages thus becomes an important task. The performance and reliability of web search engines face huge problems due to the presence of extraordinarily large amount of web data. The voluminous amount of web documents has resulted in problems for search engines leading to the fact that the search results are of less relevance to the user. In addition to this, the presence of duplicate and near duplicate web documents has created an additional overhead for the search engines critically affecting their performance. The world is becoming a single global e-village and everything is going to the web. This, in turn, make sure that all the new developments and creations happen in the web. This induces a need for a system to check data theft and fraud. For both these scenarios – **reducing search engines' burden and checking for data theft** – similarity checking systems are required. Since, duplicate detection systems cannot work on Indic languages due to their grammatical and linguistic features like inflection and agglutination, near duplicate detection is only possible, which implements approximate search. Currently, there exists no full fledged document detection system for Indic languages. Hence, this project will be a starting step towards it. Moreover, the different features used in this project like ***stemming*** are important requirements in Natural Language Processing, where they are used for other applications. This near duplicate document detection system can be used to compare the ideas generated by two news paper articles, two reports or even two books and calculate their similarity. Since Natural Language Processing is a growing field in Computer Science, the different tools used in this project are useful to Language Experts, Language Scientists and other technologists.

The approximate search algorithm and fuzzy string calculation algorithm used in this project is useful for other projects like Search Engine implementation, context based searching etc.