Near Duplicate Document Detection for Malayalam

Guided By

Mrs. Sruthy M

By:

Balasankar C

Krishnanunni R

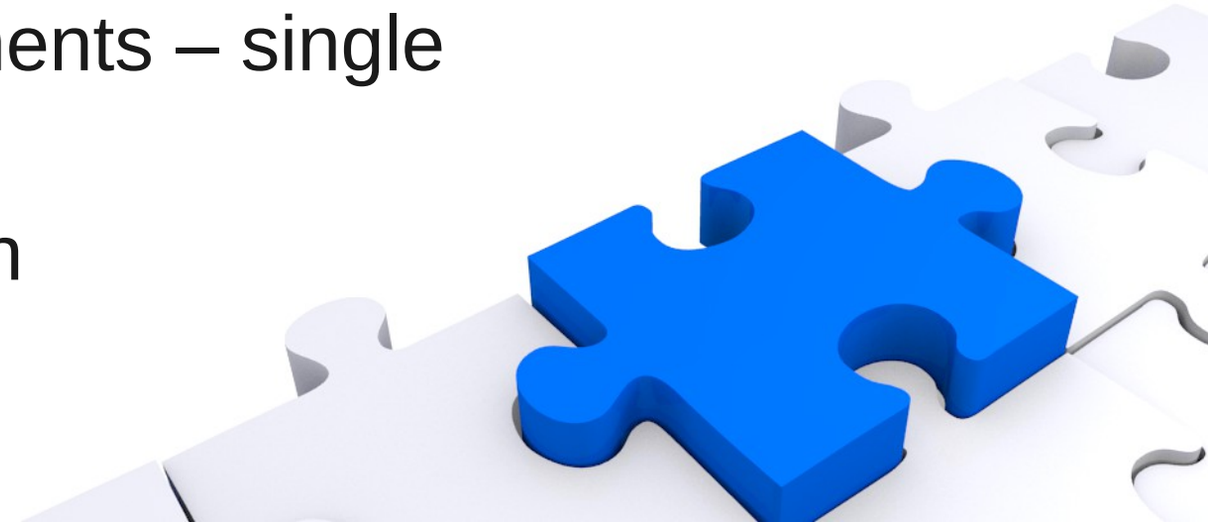
Sameena T A

Problem Defined

- Near duplicates – created by making slight variations – like insertion, deletion, replacements
- In Malayalam – additional problem of **Inflection** – words get mutated based on context.
- Eg : ഞാൻ തൃശ്ശൂരിൽ പോയിരുന്നു, ഞാൻ തൃശ്ശൂര് പോയിരുന്നു.
- Both sentences convey same idea
- When usual comparisons implemented, തൃശ്ശൂരിൽ and തൃശ്ശൂര് are different

Existing System

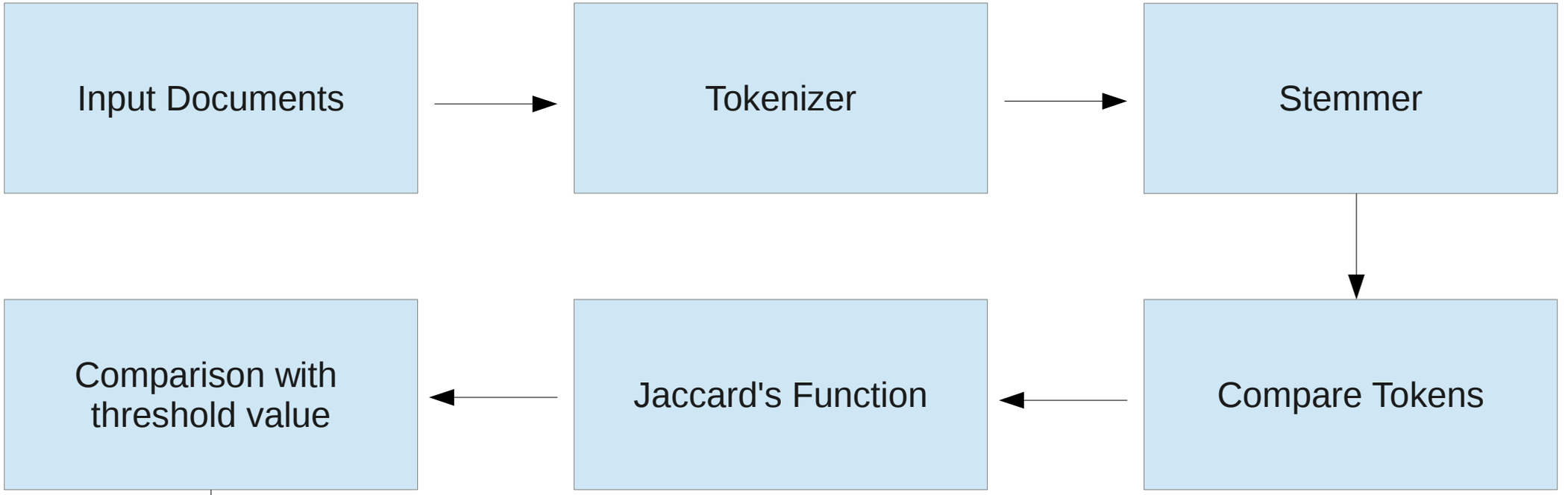
- No complete system for Indic languages
- SILPA (Swathanthra Indian Language Processing Applications project by Swathanthra Malayalam Computing – only attempt
- Doesnot handles inflections completely
- Not implemented for large documents – single sentences
- Less efficient similarity calculation



Proposed System

- Splitting documents into words (tokens)
- Tokens are stemmed words (root words)
- Stemmed word of both തൃശ്ശൂരിൽ and തൃശ്ശൂര് is തൃശ്ശൂർ
- Words arranged in ascending order to solve semantics issues
- Corresponding token sets compared using Jaccard's Similarity
- Algorithm implemented for whole document
- If similarity exceeds threshold → Near Duplicates

Flow Chart


Input Documents
Tokenizer
Stemmer
Compare Tokens
Jaccard's Function
Comparison with threshold value
Generate Output

Platform and System Requirement

- Programming Platform – Python
- GUI – PyGTK/PyQt
- Input – Text Documents
- Output – Similarity in Percentage

THANK YOU!!!