Stemming Indonesian: A confix-stripping Approach

Staff : **Mirna Adriani, Jelita Asian, Bobby Nazief, S. M.M. Tahaghoghi and Hugh E. Williams**
Students : **-**
Sponsors : **-**
Email : **mirna@cs.ui.ac.id, nazief@cs.ui.ac.id**

Stemming words to (usually) remove suffixes has applications in text search, machine translation, document summarization, and text classification. For example, English stemming reduces the words "computer," "computing," "computation," and "computability" to their common morphological root, "comput-." In text search, this permits a search for "computers" to find documents containing all words with the stem "comput-." In the Indonesian language, stemming is of crucial importance: words have prefixes, suffixes, infixes, and confixes that make matching related words difficult.
This work surveys existing techniques for stemming Indonesian words to their morphological roots, presents our novel and highly accurate CS algorithm, and explores the effectiveness of stemming in the context of general-purpose text information retrieval through ad hoc queries.

Keywords: Indonesian, information retrieval, stemming