A Faster Approach to Sort Unicode Represented Bengali Words

Abstract

Sorting Bengali words, a constituent part of Bengali language processing, Bengali data manipulation and Bengali database system comes up with a lot of challenges. A simple lexicographic ordering based on the Unicode representation does not yield the correct order of Bengali words as the character order in Unicode for Bengali differs from the order suggested by Bangla Academy. Besides, the presence of modifiers, compound characters, dual representation of some characters in Unicode as well as the precedence of vowels have made the situation even more complex. Our study aims to adapt the linguistic order for Unicode represented Bengali text while achieving maximum possible time and space efficiency. In this paper, we propose an approach to sort Bengali texts using popular algorithms with a slight modification in mapping so that it follows the linguistic order of the language and takes no extra memory. Also it shows a run time comparison with the previous works done on this topic.

References
8. The Unicode Standard 4. 0, copyright 1991-2003, Unicode, Inc.
12. Dietel, Paul. Java how to program. PHI, 2009

Index Terms

Computer Science                Artificial Intelligence

Keywords

Bengali Word Sort; Unicode Bengali sort; Bengali Linguistic Sort; Bengali Dictionary Sort; Bangla Academy Sort