Abstract

This paper describes a sorting algorithm for Bengali texts which is one of the most vital tasks for Bengali Natural Language Processing. As Unicode is much more preferable than ASCII encoding, we need to use this representation for Bengali Language. But due to some distinct properties of Bengali Language, they cannot be sorted directly using the order in Unicode character scheme. A few works have been done on this topics – some of them are for ASCII encoding whether some are for Unicode. But still they have some drawbacks and still there is no standard to sort Bengali texts. In this paper, we have discussed about the previous approaches and proposing a revised and easier procedure to sort Unicode Bengali texts. We used a mapping to simplify the sorting process. The efficiency depends on the efficiency of the sorting algorithm. This method is able to sort any Unicode Bengali texts. It will also work for Unicode text of any language if we just change the mapping part. So the process is both keyboard and language independent.
A Revised Unicode based Sorting Algorithm for Bengali Texts

11. Partha Sarathi Kar, Shantanu Mandal, Labiba Jahan, “An Improved Unicode Based Sorting Algorithm for Bengali Words”
17. Mohammad, Kazi Din: “Adhunik Bangla Byakoron O Rochona”
22. Samsad Bengali-English Dictionary
   -http://dsal.uchicago.edu/dictionaries/biswas-bengali/
23. Ishida, Richard - Bengali script noteshttp://rishida.net/scripts/bengali/

Index Terms

Computer Science Algorithms
Keywords