Abstract

Urdu is a national language of Pakistan and spoken more than 200 million people use it as a verbal and written communication. There exists a large amount of unstructured Urdu textual data in the world; by applying data mining techniques useful information can be achieved. However it seriously lacks processing capabilities to develop innovative systems based on Urdu language. In this paper, authors present a rule based stemming method for Urdu language that has the ability to cope the challenges of Urdu infix stemming. The proposed stemming method generates the stem of Urdu words by removing prefix, infix and postfix from it. In this proposed Urdu stemming technique, authors have introduced two novel classes of Urdu infix words and a new minimum word length rule. To generate stem of Urdu word that belongs to proposed Urdu infix word classes, infix stripping rules are developed. The proposed Urdu stemming technique is competent to generate the stem of borrowed words and compound words, as well. The proposed approach is evaluated on Urdu headline news datasets. This proposed approach is compared with existing state-of-the art technique (A Light Weight Urdu Stemmer) to
demonstrate the effectiveness of the proposed method. The proposed method provides 90\% to 95\% accuracy and shows significant improvements comparing to the Urdu stemming technique.

References


Index Terms
Keywords

Urdu stemming, stemming rules, infix stemming, stemming lists, Urdu infix classes