Language identification for South African Bantu languages Using Rank Order Statistics

Meluleki Dube

Department of Computer Science
University of Cape Town
South Africa

June 2018

Abstract

A lot of research has been done in the field of language identification. However, only a small proportion of these methods used for classification have been tested with the Bantu languages. In this paper we then look at one of the methods, and how it performs for the Bantu languages. The method used in this research is n-gram counting using rank orders. Using this method, we investigated how varying the testing chunk size and learning size affected the accuracy of correctly identifying the languages. The highest accuracy obtained was 99.3% with testing size of 495 characters and learning size of 600000 characters. The lowest accuracy obtained was 78.72% when the testing size was 15 characters and learning size was 200000 characters.

Keywords N-grams, N-fold cross validation, Rank Order statistics