Exploring Feature Reduction Techniques for Indic Script Text Clustering

Abstract:

Increasingly large text datasets and the high dimensionality associated with natural language especially Indic languages create a great challenge in text mining as important information is usually lost in the mining process. In this project a systematic study is conducted, in which document representation methods are used in combination with different Dimensionality Reduction Techniques, in the context of Telugu Text Clustering problem. The datasets used are of different subject matter such as Literature, Sports and Politics.

The dimensionality reduction methods include Principal Component Analysis (PCA), Kernel Principal Component Analysis (KPCA). Results are compared in terms of clustering performance and time of execution, using the k-means clustering algorithm. This entire project is to be done using python.

Group Members:

Abhay.A.Bhat (16011A0501)
Priyatosh Tripathy (16011A0521)
V. Sai Nagendra (16011A0524)

_____

Signature of Professor
(Dr. B Padmaja Rani)