Abstract

Text Mining is a field that extracts hidden, not yet discovered, useful information from the text document according to user’s query. And Text Classification is one of the text mining tasks to manage the information efficiently, by classifying the documents into classes using classification algorithms. Any text classification method uses a set of features to characterize each text document, where these features should be relevant to the task at hand. Not much work has been done for Punjabi text classification. Adequate annotated corpora are not yet available in Punjabi. This paper introduces preprocessing techniques, features selection methods for Punjabi and classification algorithm to classify the Punjabi Text documents.

References

- Nawei Chen and Dorothea Blostein, 2006, “A survey of document image classification:
Algorithm for Punjabi Text Classification


- Jiawei Han, Michelin Kamber, 2001, “Data Mining Concepts and Techniques”, Morgan Kaufmann publishers, USA, 70-181.
- Wenmin Li, Jiawei Han and Jian Pei, 2001, “CMAR: Accurate and Efficient Classification Based on Multiple Class-Association Rules”, IEEE International Conference on Data Mining - ICDM, pp. 369-376, DOI: 10.1109/ICDM.2001.989541.
- Xiaoxin Yin, Jiawei Han. CPAR, 2003,” Classification based on Predictive Association Rules”, in Proceedings of SDM, doi=10.1.1.12.7268.
- Fernando Berzal, Juan-Carlos Cubero, Nicolás Marín, Daniel Sánchez, Jose-María Serrano, Amparo Vila, “Association rule evaluation for classification purposes”.
- Dat Huynh, Dat Tran, Wanli Ma, Dharmendra Sharma, 2011, “A New Term Ranking Method Based on Relation Extraction and Graph Model for Text Classification”, Faculty of Information Sciences and Engineering, University of Canberra ACT 2601, Australia.
- Darvinder kaur, Vishal Gupta, “A survey of Named Entity Recognition in English and other Indian Languages”, Published in IJCSI International Journal of Computer Science Issues,
Algorithm for Punjabi Text Classification

- Kavi Narayana Murthy, “Automatic Categorization of Telugu News Articles”, Department of Computer and Information Sciences, University of Hyderabad, Hyderabad, doi: 202.41.85.68.
- Punjabi Corpus

Index Terms
Computer Science
Natural Language Processing
Keywords
NLP  Text mining  Text Classification  Features extraction