Abstract

In recent years, a lot of algorithms have been proposed for the classification of the documents. Most of works done have been on English language and recently there have been works on some languages such as Chinese, Arabic, etc. In some cases, there were classifications on the Persian texts which have become essays or online projects. One of the algorithms that have been used most frequently in text Classification is KNN algorithm which is more frequently in the texts Classification in the English language. In order to use these algorithms we need suitable dataset of Persian texts, which unfortunately these data are not available to Persian Texts Classification. So the our first and second phase in this project are extracting the keywords and creating Admissible dataset for the classification of the Persian texts, and The third phase of this project is to implementing a software for the classification of the Persian texts using the extracted keywords. In this essay, we have reviewed and paid attention to some challenges of searching and classifying the Persian texts, and we have also implemented an application in order to extract the admissible dataset for the classification of the Persian texts with statistical approach or with KNN and N-gram and etc, which produces some suitable and usable dataset for the classification of the Persian texts. In the last phase we have also implemented an application in order to classify the Persian texts with a statistical approach.
The Classification of Persian Texts with Statistical Approach and Extracting Keywords and Admissible Dataset

References

- Andrew Roberts, January 2009, Grammatical Inference and Corpus Linguistics, Submitted in accordance with the requirements for the degree of Master of Philosophy, The University of Leeds School of Computing.
- Shahla Nemati, Mohammad Ehsan Basiri, Persian documents classification using KNN algorithm.
- The HAMSHAHRI collection of Tehran University: www.ece.ut.ac.ir/dbrg/hamshahri
- The Website of Noor text mining group is: www.textmining.noorsoft.org

Index Terms

Computer Science Text Processing

Keywords

Persian Texts Classification Statistical Classification extracting Persian texts keywords

extracting dataset