Abstract

As the world is moving towards globalization, digitization of text has been escalating a lot and the need to organize, categorize and classify text has become obligatory. Disorganization or little categorization and sorting of text may result in dawdling response time of information retrieval. There has been the ‘curse of dimensionality’ (as termed by Bellman)[1] problem, namely the inherent sparsity of high dimensional spaces. Thus, the search for a possible presence of some unspecified structure in such a high dimensional space can be difficult. This is the task of feature reduction methods. They obtain the most relevant information from the original data and represent the information in a lower dimensionality space.

In this paper, all the applied methods on feature extraction on text categorization from the traditional bag-of-words model approach to the unconventional neural networks are discussed.
21. H. Uguz, A two-stage feature selection method for text categorization by using information gain, principal component analysis and genetic algorithm, Knowledge-Based
22. H. Uguz, A hybrid system based on information gain and principal component analysis for the classification of transcranial doppler signals, Computer Methods and Programs in Biomedicine 107 (2011) 598–609.


27. K. Pearson, On lines and planes of closest fit to systems of points in space, Philosophical Magazine 1 (1901) 559–572.

28. J. Shlens, A tutorial on principal component analysis, Systems Neurobiology Laboratory, University of California at San Diego, 2005.


35. Xiang Zhang, Junbo Zhao, Yann LeCun, Character-level Convolutional Networks for Text Classification.


Index Terms
Computer Science Information Systems

Keywords
Bag of words algorithm