A NOVEL ALGORITHM FOR CORRECTING LEXICAL ERRORS IN DATA MINING USING LEVENSHTEIN DISTANCE AND HIERARCHICAL CLUSTERING

M.TECH (CSE), GUIDE: KIRAN JYOTI

ABSTRACT

Intelligent text mining is subject that has caught up the attention of most business house and data researchers. In 2013, 5 Exabyte of data is produced on daily basis this data without further analysis and summarization is wasted. Hence researchers has developed many algorithm and systems to record, analyse, filter and summarize the produced data so that important business can be taken effectively, efficiently and in within no time. But small spelling or grammar error found in a textual data can register them as noise and thus losing important piece of information. Hence correcting those mistakes before realization is of paramount significance. But since the number of textual information is humongous, there is a lack of time critical algorithms. Internet searchers have turn into the essential method for getting to data on the Web. In any case, late studies show incorrectly spelled words are exceptionally regular in inquiries to these frameworks. At the point when clients incorrectly spell a question, the outcomes are erroneous or give uncertain data. In this work, a hierarchical clustering based lexicon correction algorithm using Levenshtein Distance for misspelling detection and correction. Hence this research work presents an algorithm for time effective corrective measure.