Text plagiarism classification using syntax based linguistic features

Adhar Partap Singh

University of Idaho

Introduction

In this paper, author proposed a mode to detect the text plagiarism by exploring syntax based linguistic characters through natural language processing (NLP). Author proposed two phase feature selection approach to increase the effectiveness by choosing suitable set of characteristics as input set to machine learning based classifiers. The results were compared with state of art approaches and the result outcome was significant.

Description

Plagiarism is an offence worldwide, when someone copies the work and idea of other person without giving him credit. Sometimes it is hard to detect plagiarism; there are so many features which plagiarism detectors unable to catch like syntax, linguistic, parts of speech, tenses and many more. Author set different sections to detect error in existing approaches by filtering and shallow syntactic characters gave effective results. The main 3 modules embrace feature extraction, feature choice and machine learning based mostly classified. Every module is intimately within the consequent sections.

The distribution of the working mechanism was shown in work-flow graph. This has three sections linguistic feature extraction, feature selection and cross-validation with machine learning. These all were further distributed in sub shells like chunk feature, parts of speech, feature value computation and many more. They evaluated there all small and large approaches with all possible conditions. Performance of the projected approach is analyzed at the start mistreatment the overall classification measures, F-measure and a pair of Accuracy. These metrics is simply computed from the complicated matrix.

Several tests were done on the basis of their realistic and artificial plagiarism in support vector machine, naïve bayes and decision tree. To evaluate the small test conditions on PSA corpus, large test conditions on PAN corpus, check the complexity and comparison with baseline systems. From the analysis and discussion, it are often noted that the projected technique with easy and borderline linguistic options sway be effective and economical in each binary and multiclass issues of plagiarism classification.

Conclusion & Future work

The result was analyzed on the bases of general and advance classification. Baseline system showed increment with dimension reduction which indicates the potency in multiple issues. Linguistic behavior of text plagiarism reflects the simulated and realistic cases in plagiarism. The variations in behavior and utility of linguistic approaches worked well to detect the plagiarism. In future, manual plagiarism will be difficult to detect form large scale of datasets. Complexity of data sets and citations increase which create a problem to explore. May simple approach will not able to work properly so, we need high performance software and detectors.