Literature Review

Text plagiarism classification using syntax based linguistic features

Author: Vani Ka, Deepa Gupta (July 8, 2017)

Adhar Partap Singh

Stat-517

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). The planned results area unit compared with the 2 progressive approaches and that they vanquish the baseline approaches considerably. This successively reflects the cogency of syntactical linguistic options in document level plagiarism. The impact of the kind and quality of plagiarism cases upon the options extracted area unit analyzed very well. 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 analysis of those dominant factors helps to urge higher insight regarding the performance variations. the degree of plagiarism embody close to copy, lightweight revisions and serious revisions moreover because the non-plagiarism cases. On the larger take a look at conditions, the evaluations area unit conducted on instances collected from information, that area unit generally categorised as realistic plagiarism cases and artificial cases. 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.

Methods

Author discussed and praised the contribution to the work done in past by other researchers, methods and steps taken to detect plagiarism by them. They talked about classification of detecting others work by corpora based system by Lancaster and Culwin (2005). How Alzahrani, (2015)presented classification that define detection ranging from global to local detection method. N-gram and Vector Space Models were choose to detect on document level works and their extensions were mainly checked on ongoing systems. IR (informational retrieval) was widely used approach with tf-idf to represent the ranking of source. Machine Learning approach in unsupervised worked as clustering to demonstrate K-mean and Fuzzy-C-Mean.

In the planned work, we tend to contribute by utilizing nominal and straightforward shallow grammar document options for plagiarism classifications. POS and chunk based mostly options once effective preprocessing and filtrations used for feature computations, to get rid of semantically irrelevant info extractions. a good two-phase feature choice approach is employed to spot nominal and best options for plagiarism classifications. one among the key factors unnoticed by the present works is that the analysis of the behavior of options with relevancy plagiarism complexities and kinds. The quality chiefly refers to completely different level of obfuscations obligatory within the text, whereas plagiarism varieties are often chiefly classified as artificial and manual cases. In artificial cases, the document plagiarism is algorithmically generated and therefore the document text might not follow correct grammatical formation and grammar rules. These aspects analyzed within the planned work exploitation example instances, that conjointly throw light-weight on the importance of knowledge bases with a lot of realistic plagiarism cases for effective evaluations. The planned evaluations ar conducted on a bigger scale, that helps to know the performance variations with differing types of plagiarism instances. Further, the comparison of planned results is completed against 2 baseline approaches.

The main 3 modules embrace feature extraction, feature choice and machine learning based mostly classification. Every module is delineated well within the succeeding sections. Features described by authors, Linguistic feature extraction: various shallow linguistic features were extracted from the suspicious and source documents at hand. POS tags and chunks are extracted using the shallow NLP techniques, POS tagging and chunking. Part of Speech (POS) features: to extract the parts of speech documents were tokenized and attached with their relative word classes using NLTK POS5 tagger. These word classes has noun, pronoun, verb, conjunction and many more POS. Lemma and lowercase used on tokens to check their dictionary they relate. Chunk features: it was used to extract stop words semantically irrelevant, to common and a lot in amount. After stop words removed, extraction started on by the chunks with length greater than or equal to 1 (>=1) and (2). The chunks with length greater than or equal to 2 (>=2). Feature value computation: this indicates document suspiciousness, Two-phase feature selection: it tells the most valuable feature to resolve the suspicious source from classifier accuracy. Cross validation, wrapper algorithm and co-relation based ranker are applied over here. Machine learning based classification: the best feature subsets was selected, classification was done using ML classifiers. Three popular and simple classifiers was evaluated in the proposed work, Naive Bayes, Support Vector Machine and Decision Tree. Support Vector Machine: supervised machine learning algorithm which models a classification problem by making it a non-probabilistic linear binary classifier. Decision Tree: non-parametric supervised learning method that is used as a predictive model by learning decision rules from inferred data. Key word based, paragraph-based and phrasal-based query searches were used while the important noun phrases were extracted for query formulation.

General and Advance Classification Metrics: F-measure and % accuracy was analysied and it computed from confusion matrix where plagiarized and non-plagairised classes presented. The true positives (number of plagiarized documents correctly classified) , false negatives (number of plagiarized documents misclassified as non-plagiarized), true negatives (number of non-plagiarized documents correctly classified as non-plagiarized) and false positives (number of non-plagiarized documents correctly classified as non-plagiarized) were obtained from matrix.

$\text{Total Population} = TP + TN + FP + FN$; $\text{Accuracy} = TP + TN/\text{Total Population}$; $F - \text{measure} = 2 * \text{recall} * \text{precision}/\text{recall} + \text{precision}$; $\text{Fall} - \text{out} = FP/TN + FP$, evaluation equation to measure general and advance classification matric.

Result

Implementation of experiment done on Weka which was open source machine learning software. Naïve Bayes, Support Vector Machine and Decision Trees with 10 fold crossvalidation was used for result.

Evaluations on smaller test conditions based on PSA corpus

It was noted that the lowest correlation score reported was 0.423 (for fADJP > = 1). Thus, α values starting from 0.4 was selected for tuning. When α =0.4, all 14 features was selected and at α=0.66, only three features (fV, fNP > = 1, fNP > = 2) are selected. α =0.52, the accuracy increases steadily and then it remains unchanged up to α =0.58. That was when α value ranges from 0.52 to 0.58, same features get selected. After α=0.58, a drop in accuracy was noted, which then increases at α =0.62. Considering decision tree classifier, without feature selection, an accuracy of 92.63% is obtained, while with the best features the accuracy improved to 97.89 %.

Evaluations on larger test conditions based on PAN corpus

Observed made on bases of Receiver Operating Characteristic analysis, the AUC is highest for Navie Bayes in both realistic and artificial cases. In the proposed approach, decision tree or Navie Bayes can be selected as best classifier. The main aim to analyze different metrics is to substantiate the correctness and authenticity of accuracy values. The analysis with both Receiver Operating Characteristic and ET reflects that the classifier performances are reliable.

Impact of plagiarism complexity and types on features: Analysis & discussions

When the type of plagiarism is unknown it was better to go with shallow features such as POS tags for plagiarism classifications at document-level. Further, semantic based techniques and deeper NLP techniques can be utilized for passage level plagiarism analysis.

Comparisons with the baseline systems

Authors did comparison between methods used by Chong et al. (2010) and Sanchez Vega et al. (2013) , +12.63 point improvement is noted in terms of accuracy and an improvement of +7.26 respectively.

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. Unavailability of the larger data bases with manual plagiarism cases is another problem and hence the creation of such data bases are also required for the proper testing of the detection approaches. May simple approach will not able to work properly so, we need high performance software and detectors.

References:

Alzahrani, S. M., Salim, N., & Palade, V. (2015). Uncovering highly obfuscated plagiarism cases using fuzzy semantic-based similarity model. Journal of King Saud University – Computer and Information Sciences, 27(3), 248–268.

Chong, M., Specia, L., & Mitkov, R. (2010). Using natural language processing for automatic plagiarism detection. In Proceedings of the 4thinternational plagiarism conference.

Lancaster, & Culwin (2005). Classifications of plagiarism detection engines. ITALICS, 4(2).

Vani, K., & Gupta, D. (2017). Text plagiarism classification using syntax based linguistic features (ISTA-2017).(pp. 448–464).

Sanchez-Vega, F., Villatoro-Tello, E., Montes-y-Gomez, M., Villasenor-Pineda, L., & Ross, P. (2013). Determining and characterizing the reused text for plagiarism detection. Expert Systems with Applications, 40(5), 1804–1813.