IJIRCCE

e-ISSN: 2320-9801 | p-ISSN: 2320-9798

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

Volume 11, Issue 8, August 2023

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 8.379

9940 572 462 | 6381 907 438 | ijircce@gmail.com | www.ijircce.com

International Journal of Innovative Research in Computer and Communication Engineering

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 |

|| Volume 11, Issue 8, August 2023 ||

| DOI: 10.15680/IJIRCCE.2023.1108016 |

New Detection Methodoly For Palgiarism Checking Using Lexical Analysis

Dr. D. J Samatha Naidu[*1], C.Sandhya[*2,]

Principal, Annamacharya PG College of Computer Studies, Rajampet, Andhra Pradesh, India[*1]

Student MCA, Annamacharya PG College of Computer Studies, Rajampet, Andhra Pradesh, India[*2]

ABSTRACT: The abstract for a plagiarism checker using the Natural Language Toolkit (NLTK) can be summarized as follows A plagiarism checker utilizing the Natural Language Toolkit (NLTK) is a software application designed to detect instances of plagiarism in text documents. The NLTK, a powerful open-source library for natural language processing, is leveraged to analyze and compare text passages to identify similarities and potential instances of plagiarism. The plagiarism checker operates by employing techniques such as tokenization, part-of-speech tagging, and n-gram analysis to break down text documents into smaller units, identify patterns, and measure the similarity between different passages. By comparing these measures against predefined thresholds or similarity algorithms, potential cases of plagiarism are flagged for further investigation. The system may incorporate additional features such as language detection, citation analysis, and integration with external databases or online sources to enhance the accuracy and reliability of the plagiarism detection process. The plagiarism checker using NLTK aims to assist in maintaining academic integrity, ensuring originality in written work, and promoting ethical writing practices. It can be utilized in educational institutions, publishing companies, research organizations, and other contexts where plagiarism detection is crucial. Careful consideration should be given to the performance, scalability, and accuracy of the plagiarism checker, as it must handle large volumes of text, maintain a low false-positive rate, and provide efficient and reliable results. Overall, the plagiarism checker employing the Natural Language Toolkit provides a valuable tool for identifying potential cases of plagiarism and supporting the preservation of academic and intellectual honesty

KEYWORDS: Plagiarism, NLP, detection methodologies, Lexical Analysis, Semantic Analysis, NLTK, LSA, PLSA, LDA

I. INTRODUCTION

Plagiarism, the act of presenting someone else's work or ideas as one's own, is a pervasive issue in academic, professional, and creative spheres. To combat this problem, the development of plagiarism detection tools has become essential. One such tool is the plagiarism checker using the Natural Language Toolkit (NLTK). The Natural Language Toolkit, or NLTK, is a widely-used open-source library for natural language processing (NLP). It provides a comprehensive set of tools and algorithms that enable the analysis, manipulation, and understanding of human language data.

The plagiarism checker utilizing NLTK aims to detect instances of plagiarism by analysing and comparing text documents. It employs various NLP techniques to break down the text into smaller units, identify linguistic patterns, and measure the similarity between different passages. By doing so, it can effectively identify cases where content has been copied or paraphrased without proper attribution. This plagiarism checker operates based on the principle that each author has a unique writing style, characterized by patterns in word usage, sentence structure, and other linguistic features. By analyzing these patterns, it becomes possible to identify instances where a document exhibits significant similarities to existing sources.

II. RELATED WORK

"Overview of Text Similarity Metrics Used in Information Retrieval" by Shubhangi D. Joshi and Sushama N. Pawar: This paper provides an overview of different text similarity metrics and techniques used in information retrieval and plagiarism detection. It explores methods such as cosine similarity, Jaccard similarity, and the use of NLTK for pre-processing and feature extraction.

"A Plagiarism Detection System Using WordNet and Cosine Similarity" by Abdul Majid and Sheikh Faisal Rashid: The authors propose a plagiarism detection system that incorporates NLTK for pre-processing and WordNet for

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 |

|| Volume 11, Issue 8, August 2023 ||

| DOI: 10.15680/IJIRCCE.2023.1108016 |

semantic analysis. The system employs cosine similarity to measure text similarity and identifies potential cases of plagiarism based on predefined thresholds. "Plagiarism Detection in Python Programming Assignments Using NLTK" by Wei Liu, Meenakshi Mishra, and Xiang Zhang: This study focuses on plagiarism detection in Python programming assignments. It utilizes NLTK for tokenization and n-gram analysis, comparing student code submissions to detect similarities. The study demonstrates the effectiveness of NLTK in detecting instances of code plagiarism.

III. LITERATURE REVIEW

Under the title "New Detection Methodology for Plagiarism Checking using Lexical Analysis Tool",

AUTHORS: Salha M. Alzahrani, NaomieSalim, and Ajith Abraham, "Understanding plagiarism linguistic patterns, textual features, and detection methods", IEEE transactions on systems, man, and cybernetics part c: application and reviews,42(2), march 2012.

Description

Plagiarism can be of many different natures, ranging from copying texts to adopting ideas, without giving credit to its originator. This paper presents a new taxonomy of plagiarism that highlights differences between literal plagiarism and intelligent plagiarism, from the plagiarist's behavioural point of view. The taxonomy supports deep understanding of different linguistic patterns in committing plagiarism, for example, changing texts into semantically equivalent but with different words and organization, shortening texts with concept generalization and specification, and adopting ideas and important contributions of others. Different textual features that characterize different plagiarism types are discussed. Systematic frameworks and methods of monolingual, extrinsic, intrinsic, and cross-lingual plagiarism detection are surveyed and correlated with plagiarism types, which are listed in the taxonomy. We conduct extensive study of state-of-the-art techniques for plagiarism detection, including character n-gram-based (CNG), vector-based (VEC), syntax-based (SYN), semantic-based (SEM), fuzzy-based (FUZZY), structural-based (STRUC), stylometric-based (STYLE), and cross-lingual techniques (CROSS). Our study corroborates that existing systems for plagiarism detection focus on copying text but fail to detect intelligent plagiarism when ideas are presented in different words.

IV. PROPOSED SYSTEM

Document Pre-processing the system pre-processes the submitted documents using NLTK. This includes tasks such as tokenization (splitting the text into words or tokens), stemming (reducing words to their base or root form), and removing stop words (commonly used words that do not carry significant meaning). Similarity Analysis the system computes the similarity between the pre-processed documents using NLTK's similarity metrics or other techniques. This can involve methods like cosine similarity, Jaccard similarity, or n-gram analysis. These metrics compare the frequency or occurrence of words, phrases, or linguistic features between the documents to determine their similarity.

Advantages

- High Accuracy
- High Efficiency

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 |

|| Volume 11, Issue 8, August 2023 ||

| DOI: 10.15680/IJIRCCE.2023.1108016 |

V. SYSTEM ARCHITECTURE

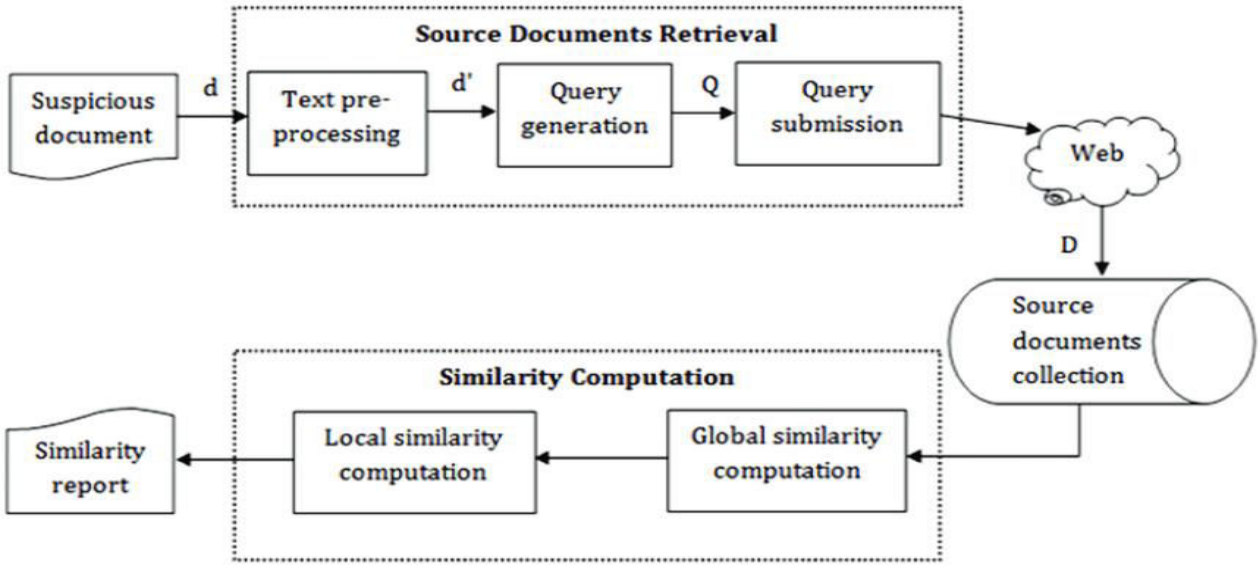


Figure1: System Architecture

VI. MODULES

- User
- System

User Module:

Use of a plagiarism checker refers to the individual or entity that interacts with the plagiarism checker tool or system. Users of a plagiarism checker can vary depending on the context in which the tool is being used. Here are some examples of different types of users for a plagiarism checker: Students, Writers, Authors.

System Module:

The "System" module is a fundamental component within the framework of the "Plagiarism Checker Using Natural Language Toolkit" application. This integral module serves as the backbone, orchestrating the seamless operation of the entire plagiarism detection system. Designed with precision and efficiency in mind, the System module encompasses a range of pivotal functions that ensure the accurate analysis and comparison of textual content.

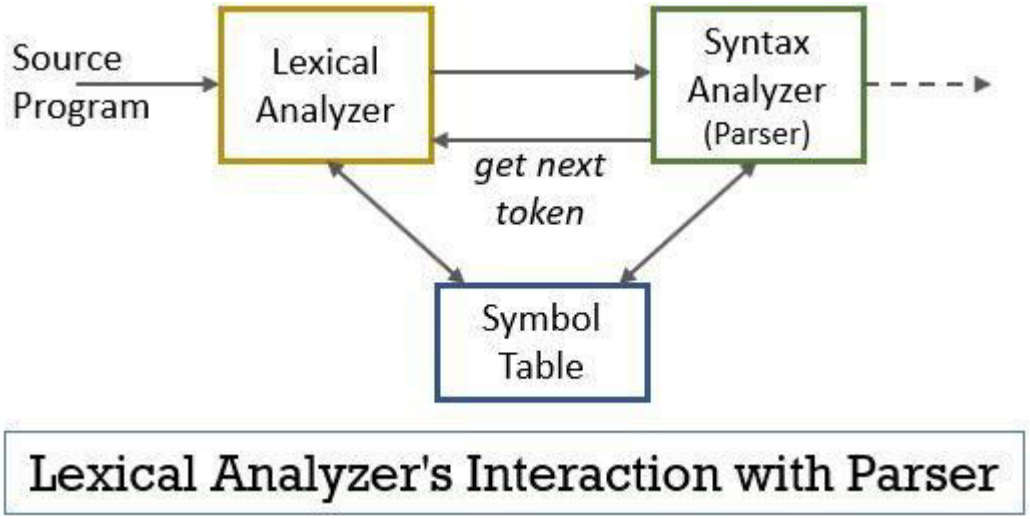
VII. ALGORITHM

Lexical analysis is the starting phase of the compiler. It gathers modified source code that is written in the form of sentences from the language preprocessor. The lexical analyzer is responsible for breaking these syntaxes into a series of tokens, by removing whitespace in the source code.

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 |

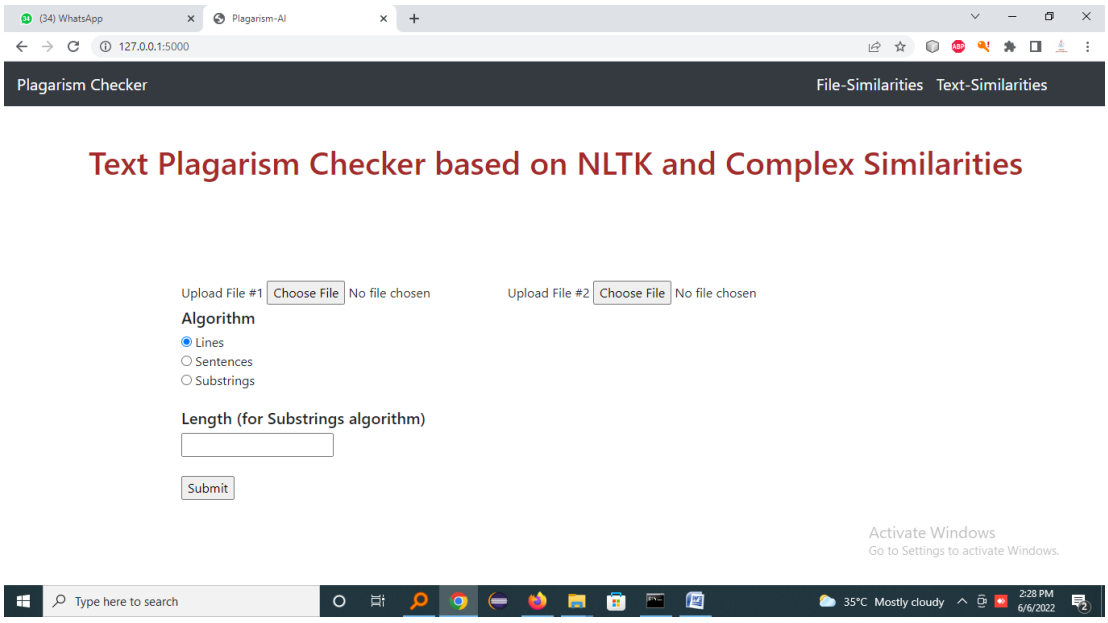
|| Volume 11, Issue 8, August 2023 ||

| DOI: 10.15680/IJIRCCE.2023.1108016 |



Lexical Analyzer's Interaction with Parser

VII. SCREEN SHOTS

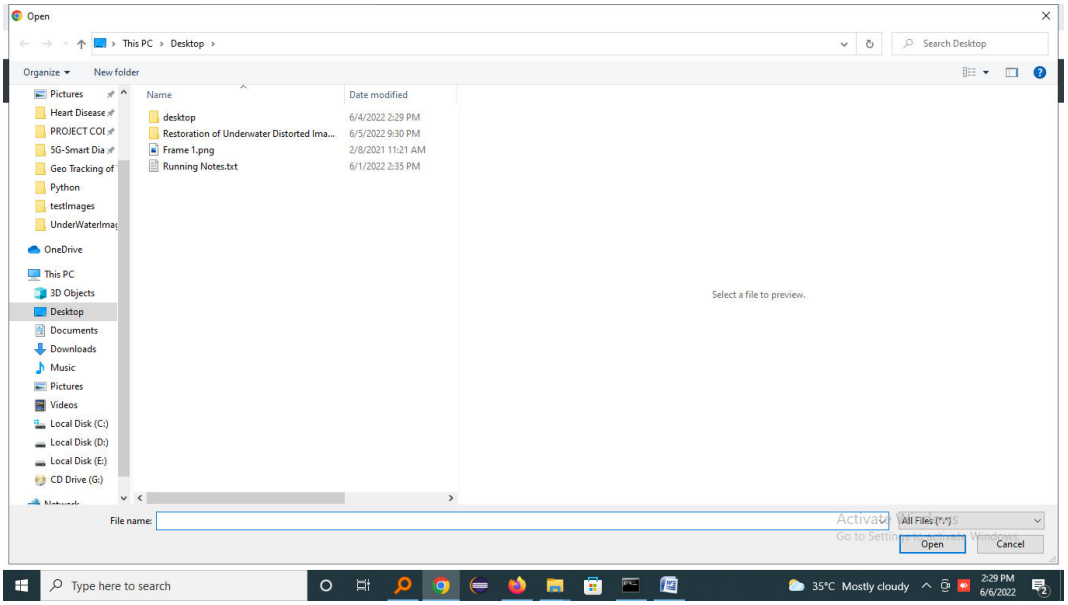


Homepage

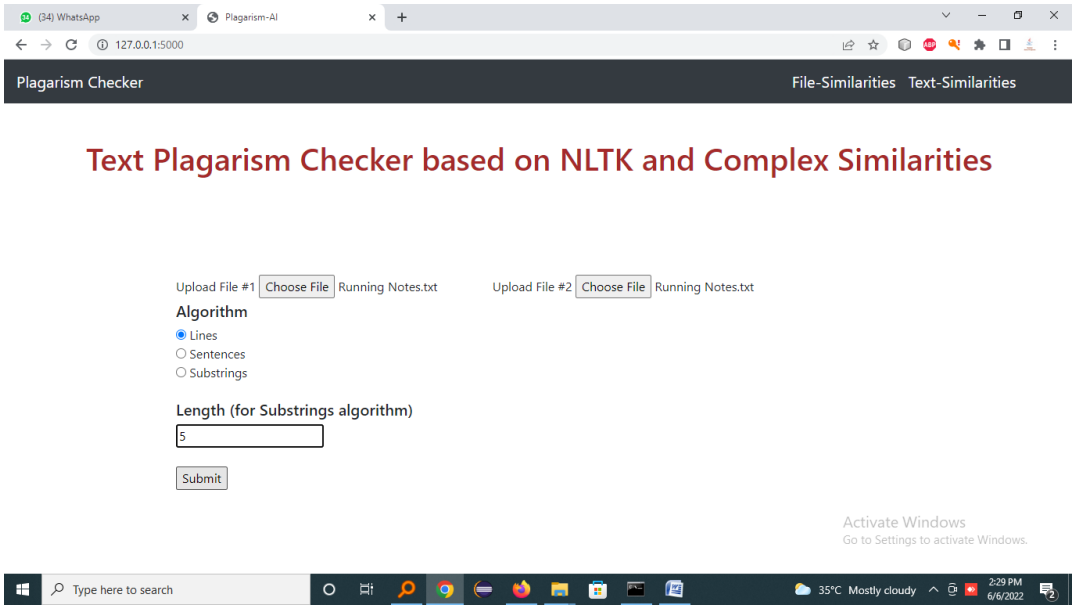
| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 |

|| Volume 11, Issue 8, August 2023 ||

| DOI: 10.15680/IJIRCCE.2023.1108016 |



UPLOAD FILES

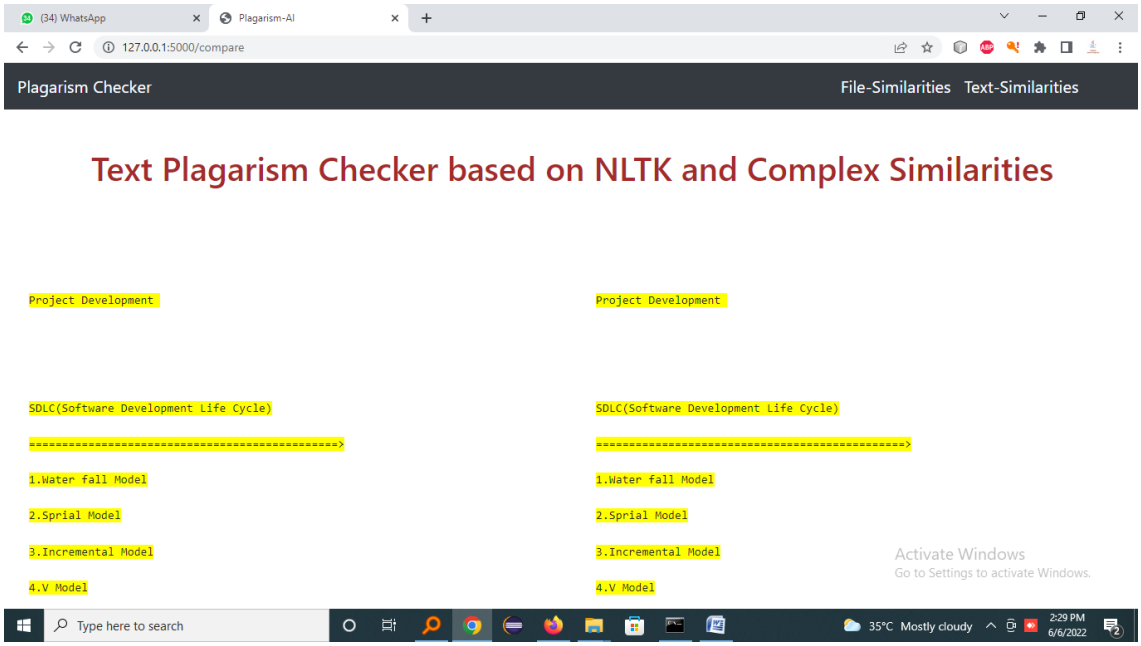


Select number of lines to check plagiarism

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 |

|| Volume 11, Issue 8, August 2023 ||

| DOI: 10.15680/IJIRCCE.2023.1108016 |



Results

VIII. CONCLUSION

This Intelligent plagiarism using fuzzy-semantic based external plagiarism detection is explored in this paper. It uses the different pre-processing methods based on NLP techniques. Here mainly lemmatization, stop word removal and POS tagging are explored. The paper provides an insight on how Ngram comparisons using POS tags can be done. It also throws light on how similarity calculation can be improved using fuzzy-semantic similarity measures. It introduces an improved fuzzy-semantic measure that can provide a significant improvement in the efficiency and accuracy of the system compared to the base method. The experimental results in terms of the PAN measures (discussed in Section 3.4) show that compared to the other methods, POSPIFS performs efficiently. According to the analysis and discussions done in Section 4, it can be observed that POS method integrated with improved fuzzy-semantic similarity measure surpass the other methods in terms of accuracy and efficiency. In future, more efficient NLP techniques can be used to improve the performance of detection system. Results can be improved by using efficient passage boundary detection conditions. Evaluation can be performed using large data sets for proper analysis and comparisons. Advanced soft computing methods and optimization techniques can be used for enhancing the system performance

REFERENCES

[1] Webster's New Collegiate Dictionary 9th ed, Springfield, Ma: Merriam 1981, pp. 870.
[2] Robert S. Nelson, Random House Compact Unabridged Dictionary: qtd. in Stepchyshyn, Vera; Library plagiarism policies.Assoc of College &Research Libraries, pp. 65. ISBN 0-8389-8416-9, 2007.
[3] Salha M. Alzahrani, NaomieSalim, and Ajith Abraham, "Understanding plagiarism linguistic patterns, textual features, and detection methods", IEEE transactions on systems, man, and cybernetics part c: application and reviews,42(2), march 2012.
[4] Ahmed Hamza Osman, Naomie Salim and Albaraa Abuobieda, "Survey of text plagiarism detection", Journal of Computer Engineering and Applications ,1(1), June 2012.
[5] Efstathios Stamatatos, "Plagiarism detection using stopword n-grams" Journal of the American Society for Information Science and Technology, 62(12), pp. 2512-2527, Wiley, 2011.
[6] Jiannan Wang, Guoliang Li, JianhuaFeng "Fast-Join: an efficient method for fuzzy token matching based string similarity join", Data Engineering (ICDE), 2011 IEEE 27th International Conference, March 2011.
[7] YuriiPalkovskii, Alexei Belov, IrynaMuzyka, "Using WordNet based semantic similarity measurement in external

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 |

|| Volume 11, Issue 8, August 2023 ||

| DOI: 10.15680/IJIRCCE.2023.1108016 |

plagiarism detection", Notebook Papers of CLEF, 2011.

[8] Ahmed Hamza Osmana NaomieSalima, Mohammed Salem Binwahlanc, RihabAlteebd and AlbaraaAbuobieda, "An improved plagiarism detection scheme based on semantic role labelling", Applied Soft Computing 12 (2012) 1493–1502.

[9] Daniel Gildea, Daniel Jurafsky, "Automatic labelling of semantic roles", computational linguistics, 28(3), 2011.

[10] Asif Ekbal, Sriparna Saha and Gaurav Choudhary, "Plagiarism detection in text using Vector Space Model, In Proc. of 12th International Conference on Hybrid Intelligent Systems (HIS), pp.366-371, Pune, 2012.

[11] Rasia Naseem and Sheena Kurian, "Extrinsic plagiarism detection in text combining VSM and fuzzy semantic similarity scheme", Journal of Advanced Computing, Engineering and application (IJACEA),2(6), December 2013.

[12] Miranda Chong, Lucia Specia and Ruslan Mitkov, "Using natural language processing for automatic plagiarism detection", 4th International Plagiarism Conference, Northrumbia University, 2010.

[13] SalhaAlzahrani, NaomieSalim, "Fuzzy semantic-based string similarity for extrinsic plagiarism detection- lab report for PAN at CLEF 2010, In Proc. of 4th International Workshop PAN-10, Padua, Italy, 2010.

[14] Martin Potthast, Benno Stein, Alberto Barron Cedeno and Paolo Rosso, "An evaluation framework for plagiarism detection", In Proc. of 23rd International Conference on Computational Linguistics, COLING 2010, Beijing, China, 2010.

INNO SPACE
SJIF Scientific Journal Impact Factor

Impact Factor: 8.379

doi cross ref

ISSN INTERNATIONAL STANDARD SERIAL NUMBER INDIA

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

9940 572 462 6381 907 438 ijircce@gmail.com

www.ijircce.com

Scan to save the contact details