Plagiarism detection tool for Indian Language documents with Special Focus on Punjabi and Hindi Language

[1]Vishal Goyal, [2]Rajeev Puri, [3]Jitesh Pubreja, [4]Jaswinder Singh
[1,3,4]Punjabi University, Patiala
[2]DAV College, Jalandhar
{vishal.pup,puri.rajeev,jitesh.pubreja,jaswinder.singh794}@gmail.com

Abstract

Plagiarism is closely linked with Intellectual Property Rights and Copyrights laws, both of which have been formed to protect the ownership of the concept. Most of the available tools for detecting plagiarism when tested with sample Punjabi text, failed to recognise the Punjabi text and the ones, which supported Punjabi text, did a simple string comparison for detecting the suspected copy-paste plagiarism, ignoring the other forms of plagiarism such as word switching, synonym replacement and sentence switching etc.

1 Introduction

The above discussed problem led to the scope of development of a specialised software that can bridge the gap. The present software tool aims at providing Successful Recognition and Reporting of Plagiarism among Punjabi and Hindi Documents. This tool is mainly aimed for Universities, Colleges or other Academic Institutions, to check the Plagiarism Report for the Submitted work, be a Ph.D. Thesis, M.Phil. Thesis or a Research Paper of some kind.

The software is built using modular approach and open source technologies. The language dependent components are kept separate from the main programme engine, so that the engine could be used with any other compatible languages. (Figure. 1).

2 Working of Software

1. The Query document is uploaded to the web based interface of the software. The uploaded document can be in plain text form, word document, pdf document or scanned document.
2. The uploaded document is first converted to Unicode format if required. The open source OCR engine Tesseract is used for converting scanned documents to Unicode format.
3. A pre-processing stage performs stop-word removal and stemming of the text, thereby reducing the size of corpus for comparison.
4. The synonym replacement module reduces the document to their base words for comparing with repository and online documents.
5. Keyword identification module identifies the important keywords from the document, that shortlists the documents having higher probability of matches.
6. The similarity matching engine uses the cosine similarity to predict the extent of match of the query document with the online as well as offline sources.

There are Three levels of Users, “Admin”, “Manager” and “User”.

The main Responsibilities for “Admin” are to Maintain and Configure the various aspects of Site for the optimal use of the System. “Admin” can also test the modular systems in isolation to pinpoint the problem if any arises during the Production Environment, which is helpful in Debugging the system down the road without putting it offline.

The “Manager” is the Organizational level head, which can add or manage users to the system under their Domain.

A “User” is the lowest form of Functional account that can be created on this site. Each User can create jobs that contains the documents, articles, papers etc. that needs to be checked.

The Plagiarism Detection tool is under the process of being trademarked in accordance with MietY, under the title “ShodhMapak”. The Copyright has been successfully granted, for the same, to Punjabi University, Patiala in conjunction with MietY.

The software tool is available online for general public at –

Proceedings of the 17th International Conference on Natural Language Processing: System Demonstrations, pages 46–47
Patna, India, December 18 - 21, 2020. ©2019 NLP Association of India (NLPAI)

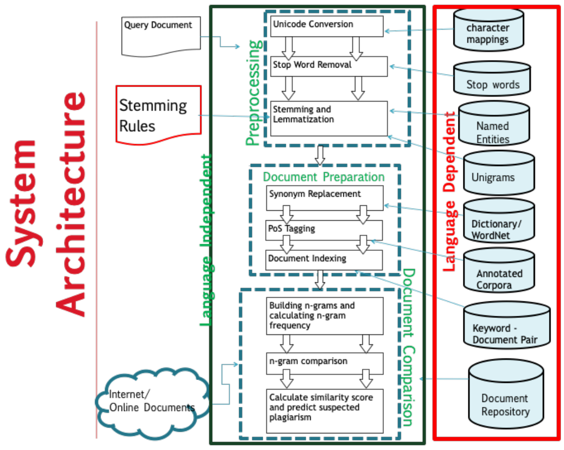

Figure. 1: System Architecture

References

R. Lukashenko, V. Graudina and J. Grundspenk, "Computer-Based Plagiarism Detection Methods and Tools: An Overview," in International Conference on Computer Systems and Technologies, 2007.

B. Martin, "Plagiarism: a misplaced emphasis," Journal of Information Ethics, vol. 3, no. 2, pp. 36-47, 1994.

H. Maurer, F. Kappe and B. Zaka, "Plagiarism - A Survey," Journal of Universal Computer Science, vol. 12, no. 8, pp. 1050-1084, 2006.

Bouville and Mathieu, "Plagiarism: Words and ideas," Journal of Science and Engineering Ethics, vol. 14, no. 3, pp. 311-322, 2008.

R. M. Karp and M. O. Rabin, "Efficient randomized pattern-matching algorithms," IBM Journal of Research and Development, vol. 31, no. 2, pp. 249-260, 1987.

D. Knuth, J. H. Morris and V. Pratt, "Fast Pattern Matching in Strings," SIAM Journal on Computing, vol. 6, no. 2, pp. 323-350, 1977.

R.S.Boyer and J.S.Moore, "A Fast String Searching Algorithm," Comm. ACM. NewYork, NY, USA: Association for Computing Machinery, vol. 20, no. 10, pp. 762-772, 1977.

R. Baeza-Yates and G. Navarro., "A faster algorithm for approximate string matching," Combinatorial Pattern Matching, vol. LNCS 1075, pp. 1-23, 1996.

"Turnitin.com User Guides," iParadigm LLC, [Online]. Available: http://guides.turnitin.com. [Accessed December 2013].

"Urkund,"Urkund,[Online].Available:http://www.urkund.com.[AccessedDecember2013].

"jPlag - Detecting Software Plagiarism," 1996. [Online]. Available: https://jplag.ipd.kit.edu. [Accessed Dec 2013].

L. Bloomfield, "wCopyFind," University of Virginia, [Online]. Available at URL: http://plagiarism.bloomfieldmedia.com /wordpress/ software/wcopyfind/. [Accessed December 2013].

"Dupli Checker - Free Online Software for Plagiarism Detection," Dupli Checker, [Online]. Available: http://www.duplichecker.com. [Accessed 22 10 2015].

W.-Y. Lin, N. Peng, C.-C. Yen and S.-d. Lin, "Online plagiarism detection through exploiting lexical, syntactic, and semantic information," in Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics, Jeju, Republic of Korea, 2012.

S. Schleimer, D. S. Wilkerson and A. Aiken, "Winnowing: local algorithms for document fingerprinting," in SIGMOD, San Diego, 2003.

S. Niezgoda and T.P.Way, "SNITCH-A Software Tool for Detecting Cut and Paste Plagiarism," in 37th SIGCSE technical symposium on Computer science education, 2006.

V. Gupta and G. S. Lehal, "Preprocessing Phase of Punjabi Language Text Summarization," Communications in Computer and Information Science, vol. 139, pp. 250-253, 2011.

V. Gupta and G. S. Lehal, "Features Selection and Weight learning for Punjabi Text Summarization," International Journal of Engineering Trends and Technology, vol. 2, no. 2, pp. 45-48, 2011.

R. Puri, R. P. S. Bedi and V. Goyal, "Plagiarism Detection in Regional Languages – Its challenges in context to Punjabi documents," Research Cell: An International Journal Of Engineering Science, vol. 5, pp. 296-304, 2011.

R. Puri, R. Bedi and V. Goyal, "Automated Stopwords Identification in Punjabi Documents," Research Cell: An International Journal of Engineering Sciences, vol. 8, no. June 2013, pp. 119- 125, 2013.

R. Puri, R. P. S. Bedi and V. Goyal, "Punjabi Stemmer using Punjabi WordNet database," Indian Journal of Science and Technology, vol. 8, no. 27, October 2015.

"Indradhanush WordNet," Dept. of Information technology, Ministry of Communication, Govt. of India.

V. Gupta and G. S. Lehal, "Automatic Keywords Extraction for Punjabi Language," IJCSI International Journal of Computer Science Issues, vol. 8, no. 5, pp. 327-331, 2011.