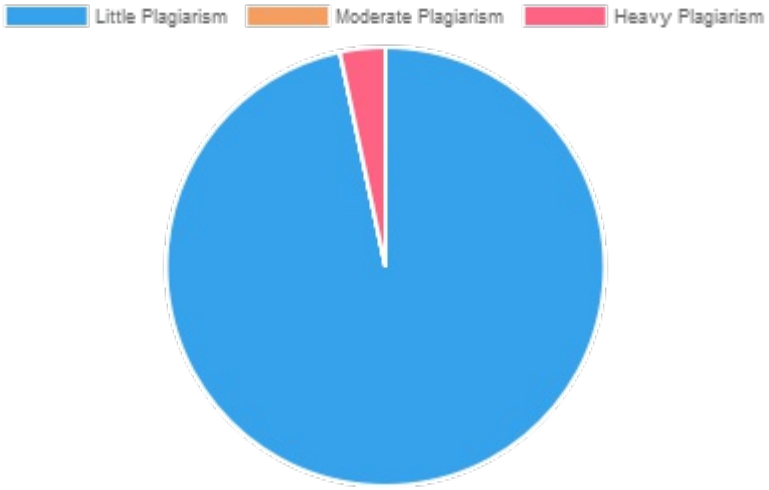
Republic of the Philippines
Bulacan State University
City of Malolos, Bulacan

Check Plagiarism Report

Checked in the System Database

University Research Office

Research Title: Author one test



Statistics Overview

Number Of Candidate Document: 30
Heavy Plagiarism: 1
Moderate Plagiarism: 0
Little Plagiarism: 29

Note: Little Plagiarism is less than 30% similarity score, 30 to 69% for Moderate and 70 to 100% for Heavy

Text Checked for Plagiarism:

See discussions, stats, and author profiles for this publication at: https://www.researchgate.net/publication/272165855 DETECTING PLAGIARISM JOURNAL WITH SHERLOCK ALGORITHM Conference Paper · September 2014 CITATIONS READS 0 524 3 authors, including: Li Za Universitas Mulawarman 39 PUBLICATIONS 31 CITATIONS SEE PROFILE Some of the authors of this publication are also working on these related projects: Distance View project 2016 FMIPA-UNMUL View project All content following this page was uploaded by Li Za on 12 February 2015. The user has requested enhancement of the downloaded file. ISBN: 978-602-9372-70-0 ARTICLES OF BALI INTERNATIONAL SEMINAR ON SCIENCE AND TECHNOLOGY (BISSTECH) II 2014 "Fundamental and Applied Research for Industrial Sustainability: Food, Agrochemical, and Information and Communication Technology (ICT)" September 2-4, 2014, BALI-INDONESIA DETECTING PLAGIARISM JOURNAL WITH SHERLOCK ALGORITHM Heliza Rahmania Hatta1, Muhammad Rasyid2, Muhamad Azhari3 Computer Science, Faculty of Mathematics and Natural Science, Mulawarman University, Samarinda, Indonesia1,2,3 Email : heliza_rahmania@yahoo.com1, rasyid37@gmail.com2, ktob34@yahoo.com3 Abstract Plagiarism is not allowed in the research journal. Therefore the similarity of research journal must be checked. Similarity is usually checked manually, so it needs long time for verification. The purpose of this research is to create a software for detecting similarity of research journal. Journal plagiarism detection software created in this research is done by implementing sherlock algorithm. Sherlock algorithm can detect document similarity by comparing similarity of each sentence inside a document with each other sentence in other documents. Sentence similarity detection is based on the same shared keyword between compared sentences. Result of test concluded that this software can detect similarity of research journal. Keywords: Journal, Sherlock, Plagiarism 1 INTRODUCTION 2 STRUCTURE OF WRITING The remainder of the paper is organized as Knowledge is growing with the expanding of information technology. This expansion provides follows : In section 3, we explain the proposed information overload on media such as research methods, observation and similarity detection journals. Implementation of information technology process. In section 4, the software testing are provide journals documentation in softcopy format. explained. Finally, in section 5, our conclusions are Journals documentation has a purpose to make the outlined. journal scans become evident in document format and can be used as research references. 3 METHODOLOGY 3.1 Sherlock Algorithm Journals that will be documented need to be checked before regarding its scientific value, which Sherlock algorithmis an algorithm for is there must be no plagiarism in it. For knowing if detecting plagiarism by comparing similarity there isplagiarism or not, there must be a check for between one sentence with other sentence. Sherlock similarity degree of journals. algorithm indicates that if there are two sentences which have different sets of keywords then these Usually, detection process is done manually, by two sentences have different content.The oppositeis reading those journals one by one. This method of if two sentences have same sets of keywords then detection is not very efficient, because it needs very these two sentences have same content. Detection long time. Therefore there should be acomputerized process is done by comparing each sentence in one similarity journal documents detection software document with each other sentence in another which can make the process for detecting similarity document [2]. of research journals becomes faster. Calculation of similarity score is done by This journal plagiarism detection software was getting the number of shared keyword from sentence implemented using Sherlock algorithm. Sherlock A that found in sentence B and then divided it with algorithm was used because this algorithm can total words in sentence A. Calculation similarity detect document similarity by comparing similarity score of sentence B is done by dividing the number of each sentence inside a document with each other of shared keywords between sentence A and B with sentence in other documents. Therefore this the total words in sentence B. Average score from algorithm is considered capable for detecting journal this calculation is the sentence similarity score. If similarity [1]. the calculation has a result of more than 80 then the compared sentences are indicated as similar [2]. According to the problem mentioned above, the main topic of this paper is to do research about Equation example of sherlock algorithm: journal similarity detection using sherlock Sentence A : Software for detecting research algorithm. This research aims to create a computerized journal similarity detection using document similarity Sherlock algorithm. 1 ISBN: 978-602-9372-70-0 ARTICLES OF BALI INTERNATIONAL SEMINAR ON SCIENCE AND TECHNOLOGY (BISSTECH) II 2014

“Fundamental and Applied Research for Industrial Sustainability: Food, Agrochemical, and Information and Communication Technology (ICT)” September 2-4, 2014, BALI-INDONESIA Sentence B : Computerized software for detecting Equation for calculating similarity percentage research document similarity by dividing number of similar sentences detected with total sentences in document. There are 3 way |(6 *100)| Ǵ |(6 *100)| for determining similarity between documents [5]: Similarity score Ť (6)(7) Ť 100 Ǵ 85,7 Ť 92,85 1. Testing result lower than 30% means those 2 2 document considered has little plagiarism. 2. Testing result between 30-70% means those There are 6 shared keywords between sentence A and sentence B which are, “software”, “for”, document considered has moderate plagiarism. “detecting”, ”research”, “document”, “similarity”. 3. Testing result more than 70% means those Insentence A there are 6 words. Insentence B there are 7 words, so the average similarity score is 92,85. document considered has heavy plagiarism. Sherlock algorithm flowchart is shown in Figure 1. 3.3 Observation Observation aimed for collecting data that needed before software implementation. Data collected were research journal documents obtained from the internet in doc and docx formats. This research used 15 journal documents as test documents and 25 journal documents as source documents. Journal documents that were collected mostly were journals with similar topic. 3.4 Similarity Detection Process Figure 1 Sherlock Algorithm Flowchart 3.2 Similarity Percentage Calculation Similarity percentage is the comparation degree of percentage similarity between tested documents, Figure 2. This similarity will give a result of a score which will be a reference for determining percentage similarity degree on a tested document. The number of similarity percentage is affected by the similarity degree from tested document. If similarity percentage is higher, then similarity degree will be higher [3]. Percentage similarity calculation between two documents using equation [4] : Similarit(y%)Ť Totaal mounotfdetectedsentences *100% Totaal mounotfsentencesindocument (1) Figure 2 Similarity Detection Flowchart 2 ISBN: 978-602-9372-70-0 ARTICLES OF BALI INTERNATIONAL SEMINAR ON SCIENCE AND TECHNOLOGY (BISSTECH) II 2014 “Fundamental and Applied Research for Industrial Sustainability: Food, Agrochemical, and Information and Communication Technology (ICT)” September 2-4, 2014, BALI-INDONESIA Firstly user input a test document as a journal Journal database form is shown in figure 7. with doc or docx format and journal’s year into User can manage data in journal database. User can preprocessing step. Preprocessing aimed to get input, edit and delet journal data. Input data is done keywords from document. Preprocessing step of a by pressing button for input data, and then software document consists of : showed form for input journal data as shown in 1. Tokenizing document into a list of sentence figure 8. using dot separator (.) Editing and deleting can be done by choosing 2. Case folding sentence by changing each capital the journal data first. if user presses edit button, software will show a form as shown on figure 8 that letter into lower case. consists of journal data which is going to be edited. 3. Filtering sentence process for stopword User can edit journal data and then press button for saving data. Journal data can bedeleted by choosing removal. data journal which is going to be deleted and then 4. Deleting duplicate word to get keywords from pressing button to delete data. sentence. The next step is similarity detection of each sentence inside document using sherlock algorithm. If the similarity score was more than 80, then the total number of detected sentences will be accumulated. Percentege similarity was calculated by dividing the number of detected sentences with the total sentences in document. 4 RESULTS AND DISCUSSIONS Figure 3 Input Journal Testing software was done by detecting journal Figure 4 Input Journal Result similarity. Testing process was done by comparing similarity between test journal document with each source journal. This aimed to test software for detecting journal similarity. Testing was done by inputting journal document. Journal can be inputted by pressing browse button. Software will show a dialog for choosing documentas seen in figure 3. Figure 3 is the dialog shown for choosing journal document. User can choose one of the journals listed and then pressing open button. Document that can be inputted are documents with doc and docx format. Figure 4 is the input journal result. Document that had been inputted automatically went through preprocessing step. The preprocessing result and preprocessing time can be seen for examination. The next step was detecting document percentage similarityby comparing with each document in the database. Document comparing step was done by pressing detection button. The next step was similarity detection by comparing test journal witheach journal in the database. If the detection process is finished, the result of journal similarity is ranked from highest percentage into lowest percentage as shown in figure 5. Figure 6 showed the detailed detection result.In detailed detection, the software detected similarity by comparing each sentence in test journal with each other sentence in journal chosen by user. If there are similar sentences found, it will be marked by red color. 3 ISBN: 978-602-9372-70-0 ARTICLES OF BALI INTERNATIONAL SEMINAR ON SCIENCE AND TECHNOLOGY (BISSTECH) II 2014 “Fundamental and Applied Research for Industrial Sustainability: Food, Agrochemical, and Information and Communication Technology (ICT)” September 2-4, 2014, BALI-INDONESIA Figure 5 Detection result Figure 8 Input Document Into Database Figure 6 Detailed Detection Result Table 1 showed performance testing of Figure 7 Journal Database software for detecting journal similarity. Detection time is the time needed for comparing similarity between test document with each source document in database. According to table 1, the average detection time of this software for comparing with 25 source documents is 13,59 second, so the average time needed for detecting one document is 0-54 second. According to table 1, document size did not affect the detection time. There were larger documents size with faster time and smaller documents size with slower time. This happened because documents had been through preprocessing step before the similarity detection. Similarity percentage in table 1 was the highest percentage picked from similarity detection result. Although some of tested documents have similar topic with source document, most of documents tested had a score below 30% percentage similarity because there are few similar senteces detected. However there is also test document with similarity percentage higher than 30% detected because of high sentence similarity. Table 1 Performance Testing Test Document Size Time Similarity (KB) (s) Percentage Test Document 1 2675 13,01 7,22 % Test Document 2 947 7,61 11,65 % Test Document 3 2336 8,45 15,72 % Test Document 4 410 7,16 54,90 % 4 ISBN: 978-602-9372-70-0 ARTICLES OF BALI INTERNATIONAL SEMINAR ON SCIENCE AND TECHNOLOGY (BISSTECH) II 2014 “Fundamental and Applied Research for Industrial Sustainability: Food, Agrochemical, and Information and Communication Technology (ICT)” September 2-4, 2014, BALI-INDONESIA Test Document 5 405 15,89 10,93 % [4] Mutiara, A.B. and Agustina, S. “Anti Test Document 6 1659 28,34 19,42 % Plagiarism Application With Algorithm Karp- Test Document 7 1194 14,43 13,84 % Rabin”. Thesis in Gunadarma University, Test Document 8 651 12,62 24,04 % Jakarta, 2008. Test Document 9 985 16,77 20,83 % Test Document 10 1167 7,69 11,81 % [5] Sastroasmoro, S. “Beberapa Catatan tentang Test Document 11 1641 13,08 9,00 % Plagiarisme”. Departemen Ilmu Kesehatan Test Document 12 751 13,46 14,80 % Anak Fakultas Kedokteran Universitas Test Document 13 577 11,15 15,03 % Indonesia, Jakarta, 2010. Test Document 14 1759 17,84 10,07 % Test Document 15 1678 16,38 15,44 % Testing results indicated t although some journals had similar topic did not mean that it had plagiarism inside. Therefore this software have detailed detection feature for identification and examination purpose to minimize this problem. 5 CONCLUSIONS According to software implementation, it was concluded that Journal plagiarism detection software using Sherlock algorithm can detect similarity of research journals with an average detecting time of 13-59 second. This software can rank journal similarity from the highest percentage into the lowest percentage. Although some journals have similar topic, it did not mean that it had high plagiarism, therefore this software can show detailed detection result by showing red marked similar sentences for easier identification. This research used 25 source documents for sample. If there are more source documents, then the chance and effectiveness for finding similar journals will be higher, but it will take slower detection time. 6 REFERENCES [1] Rasyid, M. “Pendeteksian Kemiripan Jurnal Hasil Penelitian Dengan Menggunakan Algoritma Sherlock”. Skripsi Ilmu Komputer, Universitas Mulawarman, 2014. [2] White, D. R. and Joy, M. S. “Sentence-Based Natural Language Plagiarism Detection”. ACM Journal of Education Resources, Vol 4 No. 4 pp.1-20, 2004. [3] Surahman, A. M. “Perancangan Sistem Penentuan Similarity Kode Program Pada Bahasa C Dan Pascal”. Jurnal Sistem dan Teknologi Informasi Universitas Tanjungpura, Vol 1 No. 1, 2013. 5 View publication stats

Documents Found: 30 Date Printed: March 24th 2019, 8:23 AM

#	Title	Similarity Score
1	H	100.00%
2	Testing mahaba tooooooooooooooooooooooooooooooooooooooooooooooo	6.21%
3	test 4	6.21%
4	Sample Hidden Deleted	3.73%
5	Sample course 1 research	3.73%
6	Sample for author one	3.73%
7	Test 22	3.73%
8	Testing 10	3.73%
9	sadasdasd	3.73%
10	sampleeeeeeeeeeeeeeeeee	3.73%
11	test	3.73%
12	test 10	3.73%
13	test 2	3.73%
14	test 6	3.73%
15	test test test author one	3.73%
16	Looooooooooooooooooooooooooooooong Research Title Sample	1.86%
17	Test	1.86%
18	Test 11	1.86%
19	Test mobile	1.86%
20	Testing Research CON	1.86%
21	asd	1.86%

22	author one test ulit	1.86%
23	qwe	1.86%
24	sammpoolll	1.86%
25	test 3	1.86%
26	test 7	1.86%
27	test 8	1.86%
28	test 9	1.86%
29	w	1.24%
30	test 5	0.62%
31	Sample Hidden	0.00%
32	Test 23	0.00%
- Nothing Follows -		