Student activity [output O1-A-5, en, license CC BY 4.0, 14 September 2018]

Usage of computer software for text matching

Date: 2018-09-14

Information about the use of this material:



This work is licensed under a Creative Commons Attribution 4.0 International Licence.

You are free to share, copy and redistribute the material in any medium or format. You are free to adapt, remix, transform, and build upon the material for any purpose. You must give appropriate credit, provide a link to the license, and indicate if changes were made. You may do so in any reasonable manner, but not in any way that suggests the licensor endorses you or your use.

Additional information about CC licensing: <https://creativecommons.org/licenses/by/4.0>

citation:

[original source] Kreicbergs, J. (2009). Datorprogrammu izmantošana plaģiāta kontrolē. / Usage of computer programs for plagiarism control. Available from <https://dspace.lu.lv/dspace/bitstream/handle/7/272/datorprogrammu-izmantosana-plagiata-kontrole.pdf?sequence=1&isAllowed=y>

[title] Usage of computer software for plagiarism control [Datorprogrammu izmantošana plaģiāta kontrolē]

[date] 2018-09-14

[source] <http://www.academicintegrity.eu/wp/all-materials>

[access date]

Translation: Alla Anohina-Naumeca

### Changes made

This is a translation from Latvian to English. Some parts of the original text are not included. The included text is modified by slightly changing the terminology and introducing the global context whilst keeping the national perspective.

# Usage of computer software for text matching

The academic environment is a good breeding ground for plagiarism. The teaching and research process, in addition to other activities, requires the acquisition and analysis of information generated by previous researchers. This is a normal cycle of information and knowledge acquisition, in which prior knowledge helps to generate new knowledge. Unfortunately, due to different circumstances, it seems tempting to claim credit for the achievements of others, and in this case not only does the knowledge development cycle stop, because no new knowledge is created, but also serious academic misconduct is committed.

Plagiarism in the work of first-year students is caused by their undeveloped ability to think and analyse information independently, and to evaluate honestly the results of independent thinking in other works [1]. Young students sometimes do not understand the problem of plagiarism, and instead of feeling guilty for their deeds, they consider this practice to be normal [2]. If students do not develop the ability to work independently, the university faces a situation when students are assessed despite having not learnt to understand and critically evaluate and acquire the information presented in study materials [3].

A very large amount of information is now available - which offers a lot of research opportunities, but from an anti-plagiarism point of view, the vast amount of information is very difficult to manage.

Information availability, and the rapid development of information processing, exceeds the human capability to control these processes. Plagiarism detection, based only on human memory or the ability to compare information, has become very difficult in these circumstances. Therefore, extensive and tedious information searching and examination is done by special computer software.

A basic principle of computerized text matching tools is the comparison of information. In order to carry out this process, there must be an electronic document or a set of documents serving as an original corpus, and an electronic document that is the object of checking. The object of checking is compared to the originals to find out the extent of the matching text or similarity level.

In this simple principle there are, however, a number of nuances, which will be discussed in more detail:

1. The corpus of original documents. When designing a text matching tool, it is possible to compare the test document with a corpus of other documents, which is either an original corpus or an external corpus.

a. The original corpuses are usually electronic documents, which are collected and arranged by the text matching tool. When checking students’ work, these are usually students’ work from previous years. Other documents can also be added to the set. The larger the set of original documents, the higher the probability that the match between the original and the plagiarized document will be found.

When choosing this option, it is necessary to understand upgrading possibilities and sources of the original corpus.

b. External corpuses are usually databases created by third parties, or resources on the World-Wide Web. This operation principle is common for many text matching software tools. The object of checking is compared not only to the content of global network, but also to many closed databases accessed for a fee. In this case, the corpus of original documents is huge, and this approach can be considered as very effective.

Unfortunately, for some countries, for example, Latvia, the use of external corpuses of original documents is limited. Firstly, it must be borne in mind that examination of a document means its transfer to a third party, which may give rise to legal difficulties in terms of copyright. Secondly, external corpuses can be effectively used when examining documents written in English or German. However, the number of documents available in some national (for example, Latvian) languages on the World-Wide Web and in databases is so small that use of an external corpus would be ineffective.

2. Language factor. There two factors need to be mentioned:

a. The range of some national (for example, Latvian) language information sources in comparison with information resources in English, Russian or German is very small. This often determines the choice of students and researchers to use information sources in foreign languages. The sources of information in the national language (for example, Latvian) is usually restricted to publications from local authors and students from previous years.

b. As a result, the sources of information published in foreign languages are translated. Translations, unfortunately, reduce the possibility of plagiarism detection when comparing the documents, because the computer software is unable to translate and compare the documents written in different languages. In this situation, one must rely only on the ability of a person to evaluate the document and its relevance to the author’s intellectual abilities. There is also a positive aspect of the translations: since it is very unlikely that two different people translate the same text identically, the ascertaining of such coincidence clearly implies plagiarism.

3. Formats of electronic documents. When checking plagiarism electronically, the relevance of file formats is as important as the language used in the document. In order to successfully process checking of the document, the document being examined must be in a file format where text matching tools can recognise and compare words. Some national language features (for example, the diacritics of Latvian language) set an additional requirement - language encodings of documents must be the same. The most common file formats in which electronic documents are published are the .doc format used by MS Word and the .pdf file format used by Adobe Acrobat. Note that the formats mentioned above keep text layout, charts, tables, charts, and images. If it is necessary to check the document against a large corpus of original documents, then the process speed is important. To make it faster, the MS Word .doc file format and Adobe Acrobat .pdf format are converted to the MS Notepad .txt format, which converts the document into plain text without saving the layout of the text, images, diagrams, etc. elements. As a result, the main object of research - the text - remains, but all the rest is discarded. The file takes much less space and is processed much faster.

4. Text modification and paraphrasing. It is a mistake to believe that plagiarism is practiced only by people who cannot write anything themselves. A part of plagiarists is trying to transform and disguise the work to make it more difficultly recognizable. The multiple sources of information are used in the unfair works, changing the layout of sources. Several methods are used to make the detection of plagiarism more difficult:

a. The word order in a sentence or the whole structure of a sentence is changed;

b. Text parts are paraphrased;

c. Words are replaced by synonyms;

d. Text is supplemented with words;

e. Headings are changed;

f. Removing or replacing paragraphs.

A computer software that compares texts usually detects the direct text matches only, but the program developers have foreseen more complex cases, and computer software uses different solutions to detect plagiarism for non-identical although similar text fragments.

Computer software gives a signal of finding plagiarism or suspicious text fragments. Further evaluation work is to be done by lecturers and field’s experts.

The messages prepared by the text matching software do not yet prove the fact of unfair use. They only identify the extent to which one text coincides with the other text. Suspicious cases certainly need to be checked by specialists, because the computer software is only a technical tool that cannot evaluate whether the work really is plagiarized, or the text coincidence was caused by a correctly formatted quotation from another work [4].

Sometimes, if checking performed by computer software does not produce any results, it is possible to use a simpler method. Using some web search tool, for example, Google, it is possible to search for several randomly selected text fragments on the Internet. If a source of information available on the Internet is used in a suspicious work, there is a great chance to get to the original footprints. In addition to searching for text fragments, it is possible to type separate positions from the literature list in a search window.

Information sources used:

1. Kraus, J. (2002). Rethinking plagiarism: what our students are telling us when they cheat. *Issues in Writing* 13(1), 1827-1830.
2. Bombaro, Ch. (2007). Using audience response technology to teach academic integrity. ‘The seven deadly sins of plagiarism’ at Dickinson College. *Reference Services Review*, 35(2), 296.-309.
3. Warn, J. (2006). Plagiarism software: no magic bullet! *Higher Education Research & Development* 25(2), 195.-208.
4. Evans, R. (2006). Evaluating an electronic plagiarism detection service: the importance of trust and the difficulty of providing students don’t cheat. *Active learning in higher education* 7(1), 87.-99.

## Notes to educators

This material is a translation of an original paper written by Jānis Kreicbergs and published on the web site of University of Latvia. Jānis Kreicbergs is one of the first persons in Latvia who started to develop a national unified computerized plagiarism control system. In this material, he discusses general operation principles of text-matching software and nuances of its functioning that should be taken into account when developing new or using existent text-matching software.