European Network for Academic Integrity

Tricky question on academic integrity

An expert provided an answer on a tricky question on academic integrity. It was previously published in the regular ENAI newsletter (December 10th, 2018).

What percentage of plagiarism is acceptable?

In this newsletter we bring you a possible answer on quite common question "What percentage of plagiarism is acceptable?"

The answer was prepared by prof. Debora Weber-Wulf.

This is a question that I am often asked, as I have been testing so-called plagiarism detection systems since 2004. This question shows a three-fold misunderstanding of what plagiarism detection systems can — and cannot — do.

1. You can't quantify plagiarism

Plagiarism takes many forms. The easiest to spot and document is copy & paste of complete paragraphs or pages. But then there is mosaic plagiarism, with bits and pieces taken from different sources, a word changed here or a word inserted, or a few words deleted. How do you even begin to quantify these differences? Characters in sequence, words in sequence, bag of words overlap? If sentences are just rearranged, the text is still the same content as the source, but the character sequence is quite different.

If a text is translated from another language, for example a Wikipedia article, there will be no overlap at all, but it is still a complete plagiarism. And reusing the structure or arguments of another author without reference can also be considered plagiarism.

Some systems have been observed adding percentages, something that makes no sense anyway, but even less if the individual values have been rounded up or overlap each other. Others report percentages to two (or more!) decimal places to suggest more accuracy.

Many systems don't see that text has been correctly referenced, or react to longish phrases that must of course be identical if they are proper names or citation information. These false positives can lead to false accusations of plagiarism.

It is important to keep in mind that systems can only find what they have indexed, so there will also always be false negatives, that is, plagiarized texts that go unreported.

2. Systems report different values

People tend to assume that any value must be the "true" value. But each system uses different algorithms and databases, so each system will report different sources and different amounts of plagiarism for the same text. I have seen this in testing: one test case spanned the entire spectrum of "amount of plagiarism found" from nothing to over three quarters of the text. This is why there is no absolute number to represent the amount of plagiarism. And it shows that using multiple systems can uncover additional sources.

3. Software cannot determine plagiarism, only similarity

Software is unable to understand context, it can only give a rough estimate of the similarity between texts for various definitions of similarity. But similarity is not the same as plagiarism. It is a judgement call to classify the similarities between two texts and to determine if they warrant any sort of response or sanction. These kind of judgement calls can only be done by humans — sometimes it is even necessary to discuss with others how serious they find the text overlap.

So don't let the numbers trick you into too quick a judgement. Software is a tool, not an oracle. The responsibility for its correct use lies with the teacher, not the tool.

Debora Weber-Wulf