

Test cases for plagiarism detection software

Debora Weber-Wulff

HTW Berlin, Berlin, Germany

There are many plagiarism detection systems (PDS), which claim to be effective in finding plagiarisms text that is available digitally. These systems are marketed to universities in order to help them discover students using unreferenced sources, to companies to help them find plagiarisms of their work online, and to publishers to help them discover plagiarisms before they are in print.

Many institutions in the market for such software need some way of testing the effectiveness of the systems. In particular, they want to measure how much actual plagiarism can be detected and whether or not systems can deal with the minor changes made, for example, when paraphrasing is done. Texts with clearly defined plagiarism portions can be used to assess how effective the systems are. Original matter is also needed, so that it can be determined how the systems respond to this situation.

We have tested plagiarism detection systems three times since 2004. We have developed a collection of 41 German test cases that we use to test PDS, most recently in 2008. Our corpus now includes test cases for collusion as well, and is available online as exercises for an eLearning unit about plagiarism. For 2010 English-language test cases will be developed.

Although this number and size of paper is not enough to thoroughly exercise all aspects of the various PDS, since the plagiarism amount is known and there are a class-sized number of test cases, this allows for the simulation of a real-life use case for such systems.