

Can any one help me calculate a statistical probability

Source: <http://sci.tech-archive.net/Archive/sci.stat.math/2008-03/msg00153.html>

- *From:* flame.dawn@xxxxxxxxxx
 - *Date:* Tue, 18 Mar 2008 09:54:16 -0700 (PDT)
-

Here is the question. This concerns a claim of plagiarism. There are two indexes of a similar text numbering about 750,000 words. The first index has 27,740 terms in it, while the second index has 3,500 terms in it. The authors of the first index claim that the authors of the second plagiarized their index, but it turns out the indexes are mostly different, and only a few terms are similar. Can anyone calculate what the random similarity would be, i.e., if we assume that there was no plagiarism and that index 1 (27740 terms) and index 2 (3500 terms) were independently derived, what would be the probability that some of the terms would still be identical if the text to which the indexes refer is 80%–90% similar.

.