I knew immediately as I jogged into class, a minute late, that this period was not going to be fun. I hastily sat down, frantically got out a piece of paper and furiously started jotting down bullet points. If we’re going to detect plagiarism, we need some samples to work with. Generate many files, around 10 or so, that have varying degrees of content stolen from each other. Here are some ideas: Take paragraphs from Wikipedia, your old papers you’ve written, or from any other source. Make sure there are different amounts of overlap between files. One file should be an entire copy of some of the text of another file, two should share a large paragraph, etc. Make a pair of files where one is a direct copy of the other with a bunch of synonymous words substituted throughout. Plenty of pairs of files that are totally unrelated to each other.