Catching Plagiarists

Baker Franke
The University of Chicago Laboratory Schools
Problem Motivation

• e.g. Physics 101 at a large university
• One student helping another by giving him a copy of his assignment only to have the other turn it in (or large portions of it) as his own.
Problem Statement

• Given a set of text documents, find if any have been plagiarized in full, or in part, from some other document in the set.
What’s Nifty?

• Virtually zero prep...or a lot.
• Covers *many* areas of CS1/CS2
• Real Big-Oh implications
• Accessible by/challenging to all levels of students
• No one minds it being a console application!
• Relevance and motivation
Nift[yi]est Thing of All

• This problem **REQUIRES** a computational solution - no other way to do it.
Suggest a Strategy?

• Compare n-word chunks between documents.
Nifty Parts

1. Harder text processing
2. Real data structure choices/consequences
3. “Real World” issues
Part I: Text Processing

• Processing many documents in a non-trivial way.

• Option: just compare two documents to find a degree of similarity.

• Interesting Question: what do I need to compare?

• Did someone say regular expressions?
Part II : Data Structures

• Comparing even a small set of documents requires some structures

• Many possibilities

• Real Big-Oh concerns / analysis
Part III : Real World Issues

• Large document set quickly becomes unmanageable if poor algorithm chosen.
• Memory limits
• Output representation
• Legal issues?
Resources

• Web Site
• Assignment Sheet
“Through the duration of Heathcliff’s life, he encounters many tumultuous events that affects him as a person and transforms his rage deeper into his soul, for which he is unable to escape his nature.”

--bwa93.txt