

## APCS Lab 20: Real World Problems :: Document Similarity

You now have a compendium of knowledge about programming, data structures and algorithms. We need to put it to work. In the real world it is not always clear what the best way to program something is. In whatever walk of life you eventually inhabit, you will find that you will see solutions to some kinds of problems for which you could write a little program that would most appropriately/accurately/quickly solved it. Often the solutions are not complicated. Sometimes they are. You need to have an open mind and exercise your creativity to find a solution.

Here is one such problem...

The director of Essays-R-Us has come to the software development department because she needs you to write a piece of software that can determine the degree to which two essays are similar.

She doesn't know how to solve the problem but what she wants is for the program to output some kind of a number that will allow her to determine how similar one document is to another, and that it be consistent so that if, for example, the comparison of documents A and B yields the output **234**, and the comparison of documents A and C yields the output **527**, that A and C are roughly "twice as similar" to each other as A and B.

The director knows how to use command-line tools but would prefer a GUI. On the command line she would expect to enter a command something like:

```
%> java DocComp file1.txt file2.txt
```

A GUI would be similarly simple, allowing the user to name two files and to display the output after hitting a button.

You have a week to produce something that she can use.