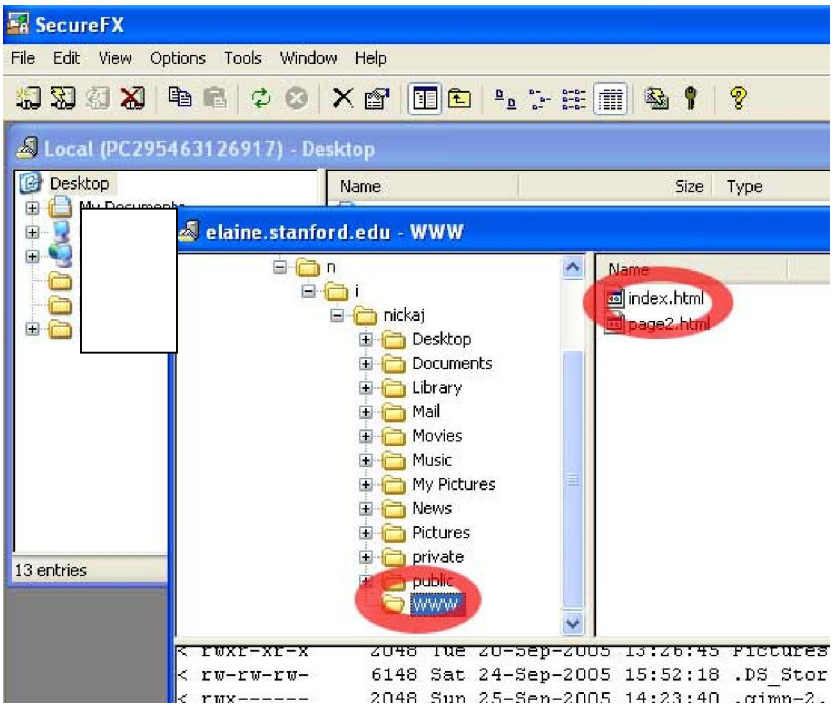
Stat 252 Homework 1

Exercise 1. Retreive your web access logs from your Leland account:

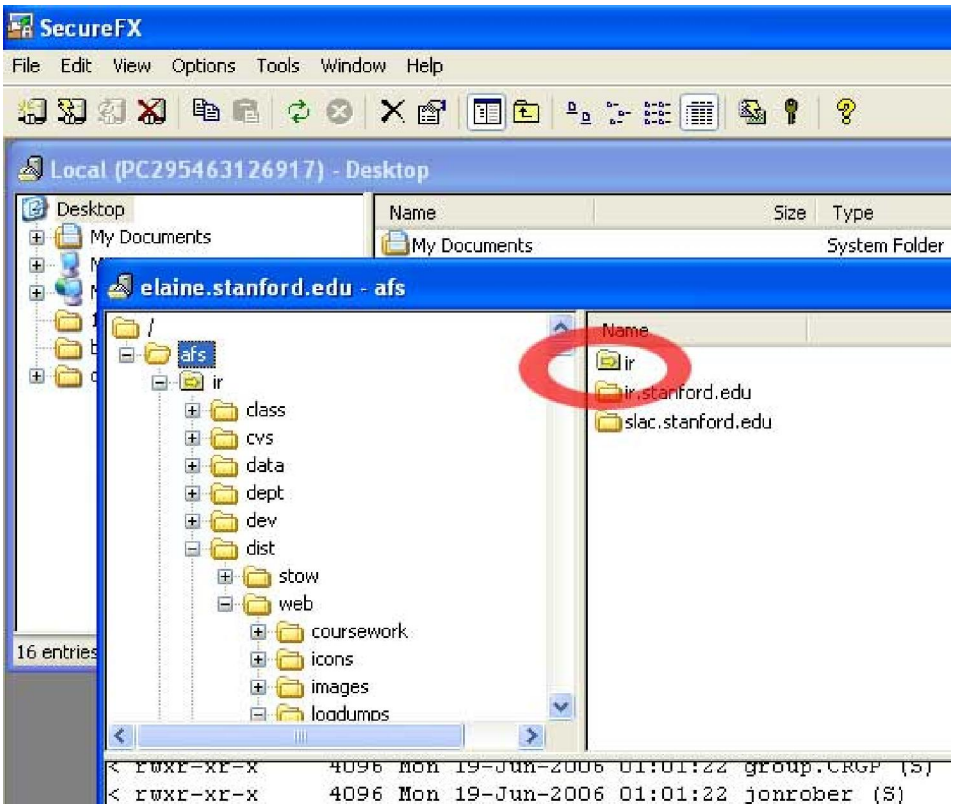
You will first need to download the necessary software:

1) PC-Leland (http://www.stanford.edu/services/ess/pc/pcleland.html) and
2) SecureFX (http://www.stanford.edu/dept/its/support/ess/pc/apps/sfx305inst.exe) from the ess.stanford.edu webpage.
3) Or if you are using a mac you can download from http://www.stanford.edu/services/ess/mac/index.html

In SecureFX connect to **elaine.stanford.edu** and log in with your SUNet ID and password. If you don't already have a webpage, you will want to transfer one to the WWW folder. The opening page should be called index.html. (See picture below)



Now you should retrieve your web access logs from the server. You can find them at
/afs/ir/dist/web/logdumps/**YOURNETID**/
You can retrieve them through SecureFX, but you will need to start from the root directory. The "ir" directory will show up as a link until you double click it and then you can traverse the rest of the path:



After creating your page and having your friends hit it a few times, you will need to wait another day for the logs to be refreshed. Comment on the format of the logs and print out a snippet.

Exercise 2: Example code for a simple webcrawler

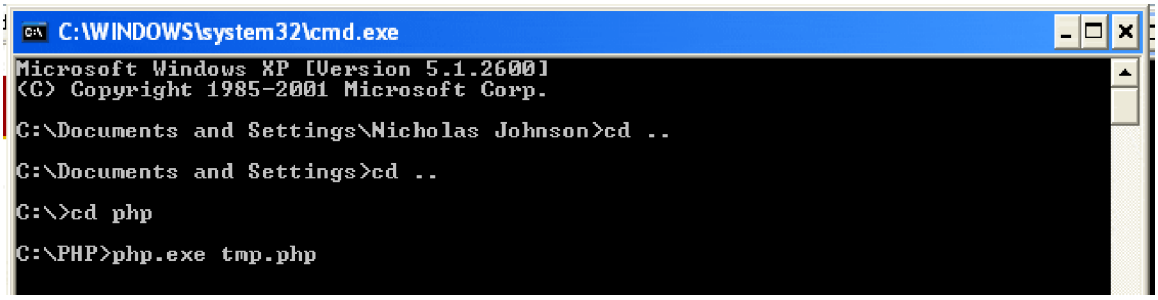
I will use PHP which can be downloaded at: http://www.php.net/downloads.php

If you're using windows use the link labeled:

PHP 5.1.4 zip package [8,919Kb] - 04 May 2006

This will give you a standalone executable which can execute .php scripts (so you don't need to configure a webserver just to run your code). These scripts take a while to load, so look for the files "php.ini-dist" and "php.ini-recommended" in the PHP folder. Open them and search for "max_execution_time". Change the value next to it from 30 (or whatever small number it was) to something large like 800. This tells PHP not to bail out of execution if your script takes too long to complete.

To execute a script, you place it in the PHP directory and call it like this:



The following code is a simple crawler written in PHP. Use it to grab the rent prices from craiglist and look for a correlation between the amount of money a person is charging and the number of words they wrote about the apartment.

Run the webcrawler (http://www.geocities.com/chsnick/MyCrawler.php.txt rename the extension to .php before running) and analyze the resulting data. Currently the script collects rent prices in the bay area. Try to get the script up and running on your computer and plot the resulting data in whatever program you feel comfortable using.

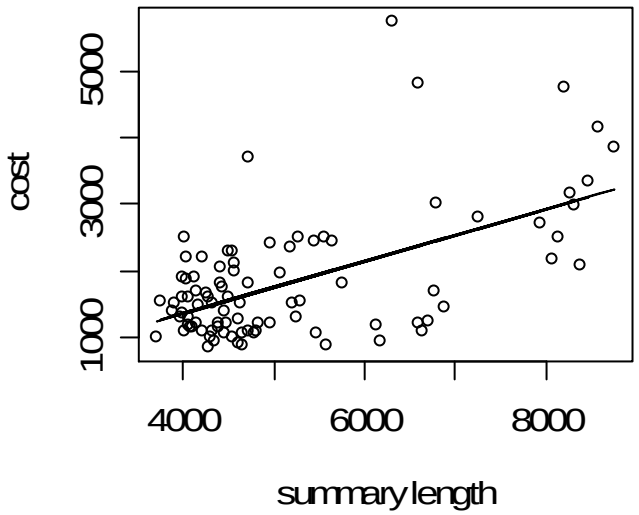
You should then modify the script to collect something else and provide some sort of plot showing what you found. You will need to adjust the parameters (how deep your search is, etc) and perhaps add new parameters to the functions to accommodate your modifications. Please provide your modified code and a link to the script on your Leland account.

Here is a snippet written in the R language (http://www.r-project.org) which fits a linear model through the data. If you would rather use or some other program to view the data, that is ok as well.

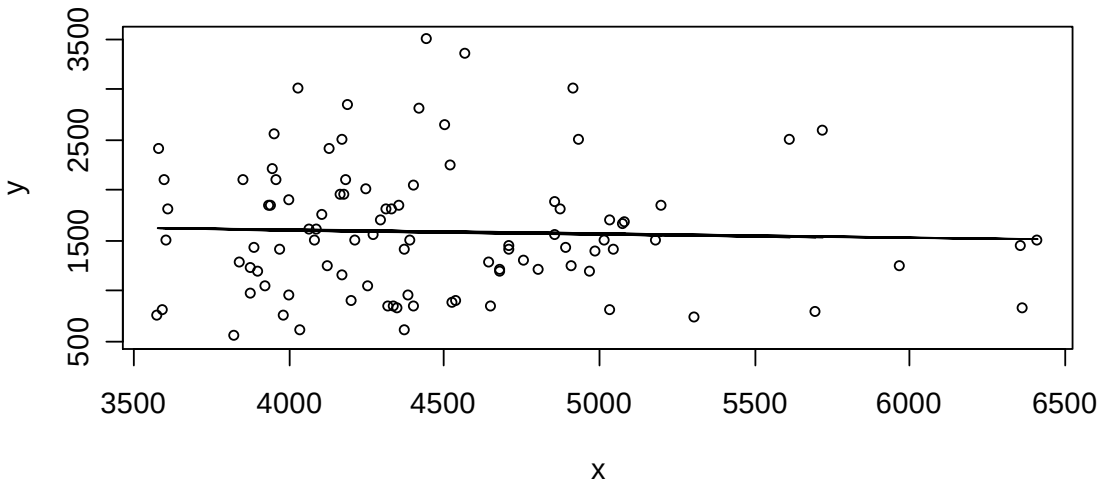
```
tmp <- read.csv(file="~/my documents/252/costs2.csv", header=T)
```

```
costs.sorted <- sort(tmp[,1], index.return=TRUE)
#there will be outliers in the collected cost data
#but the summary length is reliable.

inds1 <- costs.sorted$ix[2:(nrow(tmp)-2)]

x <- tmp[inds1, 2]
y <- tmp[inds1, 1]

mdl1 <- lm(y ~ x)
plot(x, y)
lines(x, mdl1$fitted.values)

summary(mdl1)
```

You will want to start the script at a different starting point than where it currently begins, but you also want to start it at a page with the most relevant links possible. If your code isn't careful, it may not be collecting what you think it is. When I first ran the webcrawler, I collected the following set of rent prices.



but after altering the code to search through fewer pages (no cgi links), and running it again I saw:



Finally, if you find a webcrawler written in another language, or perhaps in PHP but by someone else and you're more comfortable using it instead, feel free to. Please provide your modified code and a link to where you found the original.