Let's look at operating on directory trees, a fairly common operation when dealing with files.

It's common to want to search through a directory tree of files for matches. These days, `grep` has a built-in option for that, but let's see whether we can write that in some other (and more flexible) way.

The first thing people tend to do is look at the `find` command and see its `-exec` option; they then write something like this command. Do not use this kind of command; `-exec` is rarely the right thing to use because it is quite inefficient, because it is limited in what you can do with it, and because the syntax and quoting can get tricky.

A better way of dealing with this is the `xargs` command. It takes a partial command as its arguments, reads a list of file names on its standard input, and then applies the command to all those file names. It can do this in parallel (and there are even distributed versions of it).
In [16]: `find brown/. | xargs grep nuclear | wc`

107 3132 28944

To deal properly with file names containing spaces, you need to use one of the following two commands (look at the manual pages to see why that works). The latter is probably better behaved, since most UNIX commands expect line-oriented inputs, not null terminated inputs.

In [42]: `find brown/. -print0 | xargs -0 grep nuclear | wc`

107 3132 28944

In [45]: `find brown/. | xargs -d \n' grep nuclear | wc  # THIS REALLY SHOULD BE THE DEFAULT`

107 3132 28944

The `-l` option to `grep` tells it only to list the names of matching files. So, if we want to know the number of matching files (instead of the number of matching lines), we use this command:

In [36]: `find brown/. | xargs grep -l nuclear | sed 5q`

    brown/./cj72
    brown/./cj74
    brown/./ch21
    brown/./ch21
    brown/./cg03

In [37]: `find brown/. | xargs grep -l nuclear | wc`

35 35 455

Since the output of `find` is just a list of lines, we can apply filters to it as well, for example searching for specific file names, file name extensions, or other conditions. So, if we want to look for the term `nuclear` only in the `ch` files of the Brown corpus, we can use this command:

In [39]: `find brown/. | fgrep brown/./ch | xargs grep -l nuclear | wc`

3 3 39

We can even put another `grep` in between there to filter things:

In [41]: `find brown/. | xargs grep -l Kennedy | xargs grep -l nuclear | wc`

11 11 143

Finally, let's add our little `sed` script back in to format the output.

In [47]: `find brown/. | xargs grep -l Kennedy | xargs grep -h nuclear | sed 's/\[/^ ]+/g;s/^/s/ ' | head`

    Until Moscow resumed nuclear testing last September 1, the US and UK had released more than twice as much radiation into the atmosphere as the Russians, and the fallout from the earlier blasts is still coming down.
    On October 19, after the Soviets had detonated at least 20 nuclear devices, Ambassador Stevenson warned the UN General Assembly that this country, in "self protection", might have to resume above-ground tests.
    Now, of course, that the Russians are the nuclear villains, radiation is a nastier word than it was in the mid, when the US was testing in the atmosphere.
    After a nuclear blast, one bureaucrat suggested in those halcyon days, about all you had to do was haul out the broom and sweep off your sidewalks and roof.
    Can thermonuclear war be set off by accident? ??
    "E" stands for "execution" -- the moment a "go order" would unleash an American nuclear strike.
    Work is under way to see whether new restraining devices should be installed on all nuclear weapons.
    Only the President is permitted to authorize the use of nuclear weapons.
    The President cannot personally remove the safety devices from every nuclear trigger.
    However, the system is designed, ingeniously and hopefully, so that no one man could initiate a thermonuclear war.
    sed: couldn't flush stdout: Broken pipe