Eric Paulz

CPSC 3220-001

**Project 1**

1. Please provide command-line code and full-sentence English interpretation of the results for the following:
   1. Identify 4 image file formats and 4 source code formats present within fs1.zip and fs2.zip.
      1. *fs1:* **grep -E '\*\.jpg$|\*\.JPG$' tape\* | wc -l**
         1. *Explanation*
            1. **grep -E ‘\*\.jpg$|\*\.JPG$’ tape\*** attempts to match any file ending with the extension .jpg or .JPG within all of the tapes in fs1. The -E flag lets me extend the regular expression to include many different patterns instead of just one.
            2. **wc -l** prints the total number of lines of output, or how many files were found with a matching extension
         2. I continued to run this same command in a trial-and-error fashion with different known file extensions. I identified the following extensions:
            1. Image: .jpg, .ppm, .gif, .tiff
            2. Source: .c, .cpp, .html, .java
      2. *fs2*: same command as for fs1 except instead of tape\* being grep’s target I changed it to ls-redaction.txt.
         1. Image: .jpg, .ppm, .gif, .tiff
         2. Source: .c, .cpp, .html, .java

Note\* - I matched both lowercase and uppercase file extensions because in Linux they are considered the same, so I didn’t want them to be counted as different by grep.

* 1. How many files appear cumulatively? How many were from 1990-1995? How many from pre-1990?
     1. *fs1*: There are 2666 .jpg, 72 .ppm, 13722 .gif, 1646 .tiff, 15214 .c, 168 .cpp, 3456 .html, and 353 .java. This gives us a total of 37297 files. To identify what dates the files were from, I piped the previous grep command to an awk command before printing the word count. The command looks like this:

**grep -E '(regex to match all identified file types)' tape\* | awk '$7 >= 1990 && $7 <= 1995' | wc -l**

The awk portion of this command grabs the 7th column of output, which in this case is the year, and only returns those rows where the year is within the specified range. Running this command tells me that 36928 of the files are from 1990-1995, and only 368 of them are from before 1990.

* + 1. *fs2*: There are 578229 .jpg, 95 .ppm, 66201 .gif, 46380 .tiff, 6605 .c, 2515 .cpp, 326286 .html, and 6045 .java. This gives us a total of 1032356 files. Using the same command as I did in fs1 to identify dates (except this time the year was in column 8 instead of column 7), I found that only 7 of the files are from 1990-1995, and 228243 of them are from before 1990.
  1. What fraction of files have no registered file extension? What interpretations do you have of this?
     1. *fs1:*
        1. **cat tape\* | wc -l**
           1. Returns the total count of files in all tapes.
           2. Output is 395747
        2. **grep -E ‘\*\..\*$’ tape\* | wc -l**
           1. Returns total count of files ending in .something in all tapes.
           2. Output is 219276

Comparing these two outputs, we see that = ≈ 0.45.

So, about 45% of the files in fs1 have no registered file extension.

* + 1. *fs2:*
       1. **cat ls-redaction.txt | wc -l**
          1. Returns the total count of files in ls-redaction.txt.
          2. Output is 5760770
       2. **grep -E ‘\*\..\*$’ ls-redaction.txt | wc -l**
          1. Returns total count of files ending in .something in ls-redaction.txt.
          2. Output is 3752562

Comparing these two outputs, we see that = ≈ 0.35.

So, about 35% of the files in fs2 have no registered file extension.