**sed**

|  |  |  |
| --- | --- | --- |
| **sed** is a standard Unix utility which is a "stream editor", allowing transformations of its input. It reads the stream input line by line, performs command/script actions (if allowed on that line), and outputs the *modified* line.  Typical uses of sed:   * Execute repetitive edits to one or more files. * Convert data * Select lines containing certain data   The actions to be performed can be in either the command line directly or can be a script inside a *scriptFile*:  sed *options* '*command*' *file1 ...*  sed *options* -f *scriptFile* *file1 ...*  sed processing steps for each input line:   1. Place the line in a buffer 2. Execute the command or commands on the buffer 3. Optionally, output the buffer   Create a SedExamples folder. When logged into a fox server, please cd to the /usr/local/courses/rslavin/cs3423/sed directory and copy all the files to your sed directory.  cs1713p0.c  cs1713p0v2.c  file1  file3  file4  file5  file6  Linux  fileSSN  inventory.txt | **Example 1**: Only output to stdout all lines in cs1713p0.c that contain "printf"  $ sed -n '/printf/p' cs1713p0.c –n: not to print out from the buffer (which is everything including lines without buffers)  P for print  // about the safety of scanf and printf  printf("%-10s %-20s %10s %10s %10s %10s\n"  printf("invalid input when reading student data, only %d valid values. \n"  printf("\tdata is %s\n", szInputBuffer);  printf("%-10s %-20s %10.2f %10.2f %10.2f %10.2f\n"   * '/printf/p' means to print lines containing "printf" * The -n means to **not** output lines normally. When used with the "p", only the results of the "p" are returned.   **Example 2**: Create a new file2 from file1 replacing "Linux" with "Unix".  $ sed 's/Linux/Unix/g' file1 > file2 redirecting changes to a new file because sed put everything sto stdout without keeping anything  S:substitution pattern with linux with Unix  G: globally : all the words   * **s** means **substitute**. It is followed by a match pattern and a replacement value. * **g** means **globally** change each occurrence on the line (this doesn't mean globally in the file) * every line is sent to file2 including both modified and unmodified lines.   **Example 3**: Modify file3 removing carriage returns. For safety, make a backup of file3 in file3.sav.  $ sed -i.sav 's/\r//g' file3   * -i means to edit the file "in place" which means to modify the file. It creates a backup copy of "file3" using the specified suffix ".sav". This option is very useful if you need to modify multiple files. Change the original file (-i) and create a backup file .sav with the old file * replaces each carriage return with an empty string (i.e., deletes the carriage return)   **Example 4**: Modify file4 and file5 replacing only one occurrence of "cat" with "dog" per line for lines 1 thru 3. Save the old files with a ".cat" suffix  $ sed -i.cat '1,3 s/cat/dog/' file4 file5   * 1,3 is the range of lines to apply the substitution * The backup copies are named file4.cat and file5.cat   **Example 5**: Delete lines in file6 that begin with "#", producing file6.new  $ sed '/^#/d' file6 > file6.new ^ means beginning of the line is a specific character in this case : #   * The '/^#/d' is a pattern for deleting lines. * "^" specifies that the pattern must be at the beginning of the line. * The "d" means to delete.   **Note:** the characters after the ending slash/delimiter are called **flags**. | |
| **Using a Script File**  As stated above, the **-f** switch is used to specify a script file which allows for more complex capabilities.  Notice that example #1 printed lines which had "printf" in "//" comments. We can remove those lines.  Our script will:   * delete lines that begin with any number of spaces (including zero) followed by "//". Any number of spaces is specified by  " \*". * print lines that contain "printf" | **Example 6:** Only output to stdout the lines in cs1713p0.c that contain "printf", but don't include lines that begin (not necessarily in column 1) with "//".  # Create a script file named "example6"  $ cat >example6  The beginning of the line is a space followed by // be deleted but we have to have \ as an escape char to use / leterally  /^ \*\/\//d //delete comments line  /printf/p //print all printf comment  CTRL-D   * The first action gave a pattern that begins at the beginning of the line, followed by any number of spaces, and then followed by two slashes which had to be escaped using backslashes since slashes are our delimiters for patterns. * If the pattern in the first action is matched, the line is deleted. * The second action is only executed after the first action. If the line is deleted, the second action is not executed. * The second action prints lines containing "printf"   $ sed -n -f example6 cs1713p0.c  printf("%-10s %-20s %10s %10s %10s %10s\n"  printf("invalid input when reading student data, only %d valid values. \n"  printf("\tdata is %s\n", szInputBuffer);  printf("%-10s %-20s %10.2f %10.2f %10.2f %10.2f\n"  # Also try it on cs1713p0v2.c  $ sed -n -f example6 cs1713p0.c  // about the safety of scanf and printf  printf("%-10s %-20s %10s %10s %10s %10s\n"  printf("invalid input when reading student data, only %d valid values. \n"  printf("\tdata is %s\n", szInputBuffer);  printf("%-10s %-20s %10.2f %10.2f %10.2f %10.2f\n"  What problems do you see with that output?  Tab, print f that have multiple line |
| Example 6 only handled leading spaces. What if the "//" was preceded by tabs and/or spaces?  We need a pattern of any number of spaces or tabs in any order. This is done by "[ \t]\*" .  Note that in non-GNU versions of sed, you may have to enter a tab character instead of "\t". | **Example 7:** Only output to stdout the lines in cs1713p0.c that contain "printf", but don't include lines that begin (not necessarily in column 1) with "//". The "//" could be preceded by a combination of spaces and tabs.  # Create a script file named "example7"  $ cat >example7  /^[ \t]\*\/\//d //delete line begin with space or tab followed by //  /printf/p  CTRL-D   * The first action gave a pattern that begins at the beginning of the line, followed by any number of spaces or tabs, and then followed by two slashes which had to be escaped using backslashes since slashes are our delimiters for patterns. * If the pattern in the first action is matched, the line is deleted. * The second action is only executed after the first action. If the line is deleted, the second action is not executed. * The second action prints lines containing "printf"   $ sed -n -f example7 cs1713p0v2.c  printf("%-10s %-20s %10s %10s %10s %10s\n"  printf("invalid input when reading student data, only %d valid values. \n"  printf("\tdata is %s\n", szInputBuffer);  printf("%-10s %-20s %10.2f %10.2f %10.2f %10.2f\n" |
| Example 6 and example 7 didn't print the entire **printf** statements; instead, those examples just printed the line that contained "printf". How can we print the entire printf statement?  We need a range of lines that begin with a "printf(" and end with its ");" | **Example 8:** Only output to stdout the lines in cs1713p0.c that contain actual "printf" statements including all lines until the ");".  # Create a script file named "example8"  $ cat >example8  /^[ \t]\*\/\//d  /printf(/,/);/p specify pattern to print a complete printf  Pattern 1: printf(  Pattern 2: ,  Pattern 3: )   * The second action lists a range of lines that begin with a line containing "printf(" and ends with a line containing ");". Those are two different patterns which are in a range since there is a comma between them.   $ sed -n -f example8 cs1713p0.c  printf("%-10s %-20s %10s %10s %10s %10s\n"  , "ID", "Name", "Exam 1", "Exam 2", "Final", "Average");  printf("invalid input when reading student data, only %d valid values. \n"  , iScanfCnt);  printf("\tdata is %s\n", szInputBuffer);  return ERR\_BAD\_INPUT;  }  dAverage = (student.dExam1 + student.dExam2 + student.dFinalExam) / 3;  printf("%-10s %-20s %10.2f %10.2f %10.2f %10.2f\n"  , student.szStudentIdNr  , student.szStudentFullNm  , student.dExam1  , student.dExam2  , student.dFinalExam  , dAverage);  Notice that the result is incorrect. Why?  Because it looks for pattern in the second pattern on different line |
| We can fix the problem with example 8 as shown to the right. | **Example 9:** Only output to stdout the lines in cs1713p0.c that contain actual "printf" statements including all lines until the ");". This fixes the problem in #8  # Create a script file named "example9"  $ cat >example9  /^[ \t]\*\/\//d  /printf(.\*);/p only single line printf, \* is a quantifier  /printf(.\*);/d look for the exact same thing and delete it so they don’t reach the part that they mess up  /printf(/,/);/p print the multi line   * The second action matches a statement that begins with "printf(" followed by any characters and then has a ");" all on the same line. It then prints it. * The third action does the same pattern match as the second action, but it deletes the line; therefore, the fourth action is only executed when it doesn't match.   $ sed -n -f example9 cs1713p0.c  printf("%-10s %-20s %10s %10s %10s %10s\n"  , "ID", "Name", "Exam 1", "Exam 2", "Final", "Average");  printf("invalid input when reading student data, only %d valid values. \n"  , iScanfCnt);  printf("\tdata is %s\n", szInputBuffer);  printf("%-10s %-20s %10.2f %10.2f %10.2f %10.2f\n"  , student.szStudentIdNr  , student.szStudentFullNm  , student.dExam1  , student.dExam2  , student.dFinalExam  , dAverage); |
| Some important pattern symbols for sed **regular expressions**:  . matches any single character including newline  ^ the pattern that follows must begin at the first character  $ the pattern that precedes the $ must match at the end of the line  \* matches zero or more of the preceding character, group or bracketed list  [*list*] matches one character to any of the characters **listed** within the brackets. For convenience, range abbreviations can be used (e.g., [a-f], [0-9])  [^*list*] matches one character if it is **not** **listed** within the brackets. ^ in this case is negate .  \(*regexp*\) groups the inner *regexp*  \+ similar to \*, but matches one or more  \{*i\*} matches *i* occurrences of the preceding character, group or bracketed list  Symbols representing characters to match (e.g., ., [list], etc) are called **classes.** Symbols which dictate the amount of a preceding class (e.g., \*, +, etc) are called **quantifiers**.  Groups can be referenced later with "\1", "\2", etc.  To reference the characters ".", "^". "$", "\*", "[", "]", in a regular expression, it must be escaped with a backslash.  To reference the characters "+", "{", "}", "(", ")", just use each normally. The special meaning uses a backslash when referenced in a regular expression. | **Example patterns for sed regular expressions:**  /^The/ Matches any line that begins with the word "The"  /[a-z]\{3\}[0-9]\{3\}/ Matches a lowercase abc123 ID.  That number 3 said that it have to be 3 of them  /[0-9]\*$/ Matches any line that ends with zero or more digits. Why is that pattern not very useful? Because it will match literally anyline because we can have 0 digit char at the end of the string What would make it have to end in digits? /[0-9]\+$/  /\(bark\)\+/ Matches one or more occurrences of "bark" anything that have the word bark in it  /(.\*)/ Matches a left parenthesis followed by any number of any characters followed by a right parenthesis.  What does this pattern mean?  /[^abcd]\.$/ anything that is not abcd immediately follow by a period at the end of the line  What would be the pattern for matching lines that are empty or contain any number of spaces and/or tabs?  /^[ \t]\*$/ spaces or tab and consist of many number of them for a whole line |
| Note that sed also supports **extended regular expressions** if the **-E** or **-r** switch is specified. This changes some of the above symbols and adds some additional symbols.  . matches any character including newline  ^ the pattern that follows must begin at the first character  $ the pattern that precedes the $ must match at the end of the line  \* matches zero or more of the preceding character, group or bracketed list  [*list*] matches one character to any of the characters **listed** within the brackets. For convenience, range abbreviations can be used (e.g., [a-f], [0-9])  [^*list*] matches one character if it is **not** **listed** within the brackets.  (*regexp*) groups the inner *regexp*  + similar to \*, but matches one or more  {*i*} matches *i* occurrences of the preceding character, group or bracketed list  *pat1*|*pat2* matches either *pat1* or *pat2*  To reference the characters ".", "^". "$", "\*", "[", "]", "+", "{", "}", "(", ")", "|" in an extended regular expression, it must be escaped with a backslash. | **Example patterns for sed extended regular expressions:**  /[a-z]{3}[0-9]{3}/ Matches a lowercase abc123 ID.  /(bark)+/ Matches one or more occurrences of "bark"  /\(.\*\)/ Matches a left parenthesis followed by any number of any characters followed by a right parenthesis.  /(abc)|(xyz)/ Matches exactly "abc" or "xyz".  Using sed extended regular expressions, what would be the pattern to match a phone any phone number like (210)456-1234, but with any digits?  sed –E ‘/\([0-9]{3}\)[0-9]{3}-[0-9]{4}/’ file |
| **Exercise #1:** For privacy reasons, we want to replace Social Security Number references in emails from customers with XXX-XX-XXXX when we archive the email.  Assume the input file is fileSSN, produce an output file with SSNs replaced as described above. | sed –i.backup –E ‘s/[0-9]{3}-[0-9]{2}-[0-9]{4}/XXX-XX-XXXX/g’ fileSSN  //do it in place (-i)or back up (-i.backup) s for substitution g is for global for doing the action for every occurrence in a line: without a g we only do it 1 time |
| **Negation**  We can negate a pattern by following the pattern with a bang. | **Example 10**: Delete all lines in the file that **don't** contain "linux" or "Linux"  $ sed '/[Ll]inux/!d' Linux //! Is negate of the deletion in this case  Although Linux is well known today, Linux was not known  was a high school English teacher, if she knew Linux. |
| **Append and Insert**  **a** Append inserts one or more lines of text **after** the matched lines.  **i** Inserts one or more lines of text **before** the matched lines. | **Example 11:** We will append the Line "hello there" to each of the first three lines of the data from file5.  # create a script file named "app11"  $ cat > app1  1,3 a\ //only on line 1and 3 we do this appending action \for escape the newline character (basically an enter key)  hello there  you can also do this on 1 line: 1,3 a\\n hello there    CTRL-D  $ sed -f app11 file5  The weather was getting very bad. It was raining cats and dogs.  hello there  It was so bad that I stepped in a poodle.  hello there  what a catastrophe.  hello there  I was so tired that I needed a cat nap.  **Example 12:** Insert the two lines shown below before any lines containing "cat" or "dog" in file5.  # create a script file name "ins12"  $ cat > ins12  /(cat)|(dog)/i\ looking for line with cat and dog we do insertion  Who is watching\  a cat or dog video?  CTRL-D  $ sed -E -f ins12 file5  Who is watching  a cat or dog video?  The weather was getting very bad. It was raining cats and dogs.  It was so bad that I stepped in a poodle.  Who is watching  a cat or dog video?  what a catastrophe.  Who is watching  a cat or dog video?  I was so tired that I needed a cat nap. |
| **Change**  **c** changes one or more **lines** replacing them with the specified text | **Example 13**: Replace lines 2-6 with "eunuchs" for the linux file  $ sed '2,6c eunuchs' Linux  Although Linux is well known today, Linux was not known  eunuchs  a word that was a homonym, she was unnerved.  **Example 14**: Completely replace lines containing "Linux" or "Unix" with simply "eunuchs"  $ sed -E '/(Linux)|(Unix)/c eunuchs' Linux  eunuchs  outside of technical circles in the 1980s. I worked  for a company that was hiring technical writers in  the mid 1980s. An interviewer asked a candidate, who  eunuchs  Having never heard of the operating system, but knowing  a word that was a homonym, she was unnerved. |
| **Substitute**  We have already seen several examples for substitute (examples 2, 3, and 4 and exercise #1).  s/*matchPattern*/*replaceValue*/*flags*   * *matchPattern* is the pattern to find in a line * *replaceValue* is the value used as a replacement for the matched value * *flags:*   g - globally replacement all occurrences in the line  *i* - this is a number; replace the *i*th occurrence only  Although our examples have shown slash ("/") as the delimiter between the *matchPattern, replaceValue, and flags,* you can specify a different delimiter after the "s". This can make it easier when you need slash in your pattern or replacement value. Substitute simply uses the first character after the "s" as the delimiter  Instead of s/\/\/.\*$// substitute comments with sumthing  you could use s|//.\*$||we now have the power to do the whole block instead of having to slash all the time- this only works with substitution  ex: /\/\.\*$/d : delete the whole line in stead of replacing part of the line | **Example 15**: Examine the file inventory.txt. It contains many inventory records. The values are separated by spaces except the last data value is a product name which can contain multiple spaces and follows the phrase "Name:". We want to keep the product ID and unit price. This means that we want to remove the second, third, and fifth logical columns. Note that product IDs end in three numbers.  $ cat >example15  s/Name:.\*$// replace what ever after name up to dollar sign  s/ [0-9]+ [0-9]+//  CTRL-D  $ sed -f example15 inventory.txt  PPF001 9.95  SBB001 14.95  SBG002 14.95  BOM001 29.95  MCW001 12.45  TTP001 9.95  NHC001 9.95  SSX001 29.95    Can we use different quantifiers? ?? |
| **Alternative to Scripts when needing multiple commands**  The **-e** (lowercase) command argument can be used to provide multiple edits to a file without requiring a script file.  This is very useful when you need variable values for different sed script steps since sed doesn't provide an easy way to pass variables as parameters. | **Example 15-2:** use -e command arguments to solve example 15.  $ sed '-e s/Name:.\*$// ' '-e s/ [0-9]+ [0-9]+//' \  ? '-e ??' inventory.txt ? is just like $ at the beginning as a prompt  PPF001 9.95  SBB001 14.95  SBG002 14.95  BOM001 29.95  MCW001 12.45  TTP001 9.95  NHC001 9.95  SSX001 29.95 |
| **Next**  The next command can be used to skip lines without printing them.   * n – Skips to the next line of input and overwrites the current line. Therefore, the current line and all changes made to it will not be printed. * N – Reads the next line of input and appends it (with a newline character preceding) to the end of the current line.   sed uses newline characters (\n) to separate one line from another and when a line is copied into the pattern space the newline character is removed. This means that using the commands we've seen so far, we would be unable to remove newline characters from a file. One solution to this is the N command.  Note that when printing a line of output sed prints a newline character provided there is at least one character in the buffer. Because of these restrictions sed is not the recommended tool for removing newline characters. Instead use 'tr' | **Example 16**: Remove newline characters from any line containing 'Unix' and replace them with a tab character  $ cat >example16  /(Linux)|(Unix)/ s/\n/\t/  CTRL-D  $ sed -E -f example16 < file1  Although Unix is well known today, Unix was not known  outside of technical circles in the 1980s. I worked  for a company that was hiring technical writers in  the mid 1980s. An interviewer asked a candidate, who  was a high school English teacher, if she knew Unix.  Having never heard of the operating system, but knowing  a word that was a homonym, she was unnerved.  The above script fails to find and replace the newline characters because they have already been removed by sed before pattern matching happens.  $ cat >example16  /(Linux)|(Unix)/N  s/\n/\t/  CTRL-D  $ sed -E -f example16 < file1  Although Unix is well known today, Unix was not known outside of technical circles in the 1980s. I worked  for a company that was hiring technical writers in  the mid 1980s. An interviewer asked a candidate, who  was a high school English teacher, if she knew Unix. Having never heard of the operating system, but knowing  a word that was a homonym, she was unnerved. |
| **Input Buffers**  The sed editor provides 2 buffers to use while editing a file. These are referred to as the **pattern space** and the **hold space.**  All commands we have used so far operate on the patternspace.  Immediately after finishing all commands in the file and before looping back, sed outputs the contents of the pattern space followed by a newline unless the -n flag is set. **This is why all lines print by default.**  The hold space is an additional buffer that can be used to hold a value across multiple lines. It remains unchanged until operated on by one of the commands on the right. | The **pattern space** and **hold space** buffers have several commands that can be used to manipulate them.   * **g** – Copies contents of hold space into pattern space * **G** – Appends contents of hold space onto pattern space (Adding a newline between) * **h** – Copies contents of pattern space into hold space * **H** – Appends contents of pattern space into hold space (Adding a newline between) * **x** – Swaps values stored in pattern space and hold space |
| **Move a line**  Example 16 first uses the **h** command to store the contents of line containing the pattern 'last'. Next the **d** command is used to remove it from the pattern space so that it is not printed. Finally, on the last line of input the **G** command is used to append the line stored in the hold space onto the end of the pattern space before printing.  What would happen if there were multiple lines in the file containing the pattern 'last'? | **Example 17**: Examine the file hold\_example.txt. It has a problem; the last line is in the wrong positon and needs to be moved to the end. This can only be done if we are somehow able to save off the first line and only print it when we reach the last line.  $ cat >example17  /last/h //throw the line with the word last into the hold space  //if we want to copy multiple line with the word last use H  /last/d //delete the line  $ G $ means the last line and take the hold space and put it to the last line with \n in between . if use g instead of G then we replace the last line  CTRL-D  $ sed -f example17 hold\_example.txt  First line  Second line here  Here is the third line  Almost at the final line.  This should be the last line. |
| **Warning:** some implementations of sed restrict the max size stored in both the pattern space and hold space to 4000 bytes. GNU sed has no max buffer size provided it can malloc more memory. |

**©2017 Larry W. Clark and Rocky Slavin,** UTSA CS students may make copies for their personal use

Important example: grouping can be reference later with \1 \2 for each group

/hello / a\

World

This is equivalent to s/(^.\* hello.\*$)/ \1 \n world/ \1 represent for the whole group inside ()

Explain group 1 start with anything can be as long as possible as long as you find the word hello and after that followed by many character until the end of the line