

Evolutionary Genetics

LV 25600-01 | Lecture with exercises | 4KP

Regular Expressions

Sample	Lane	Cell	Archive
JL-2-76	S7	L001	R2_001.fastq.gz
Sample	Read	Format	

`JL-2-76_S7_L001_R2_001.fastq.gz`



`JL-2-76_R2.fq.gz`

```
mv JL-2-76_S7_L001_R2_001.fastq.gz JL-2-76_R2.fq.gz
```

move

stdin

stdout

```
z_tmp/JL-2-76_S7_L001_R2_001.fastq.gz
z_tmp/JL-2-86_S18_L001_R1_001.fastq.gz
z_tmp/JL-2-86_S18_L001_R2_001.fastq.gz
z_tmp/JL-2-96_S34_L001_R1_001.fastq.gz
z_tmp/JL-2-96_S34_L001_R2_001.fastq.gz
z_tmp/JL-3-53_S40_L001_R1_001.fastq.gz
z_tmp/JL-3-53_S40_L001_R2_001.fastq.gz
z_tmp/JL-3-63_S8_L001_R1_001.fastq.gz
z_tmp/JL-3-63_S8_L001_R2_001.fastq.gz
z_tmp/JL-3-73_S19_L001_R1_001.fastq.gz
z_tmp/JL-3-73_S19_L001_R2_001.fastq.gz
z_tmp/JL-4-20_S35_L001_R1_001.fastq.gz
z_tmp/JL-4-20_S35_L001_R2_001.fastq.gz
z_tmp/JL-4-23_S41_L001_R1_001.fastq.gz
z_tmp/JL-4-23_S41_L001_R2_001.fastq.gz
z_tmp/JL-4-33_S9_L001_R1_001.fastq.gz
z_tmp/JL-4-33_S9_L001_R2_001.fastq.gz
z_tmp/JL-4-50_S20_L001_R1_001.fastq.gz
z_tmp/JL-4-50_S20_L001_R2_001.fastq.gz
...
```

JL-2-76_S7_L001_R2_001.fastq.gz

```
z_tmp\/(\w+-\w+-\w+)_S\d+_L001_(R[12])_001.fastq.gz
```

```
mv z_tmp/$1_S*_L001_$2_001.fastq.gz a_data/gz/\$1_$2.fq.gz
```

```
#!/bin/bash
```

```
## Rename Sequencing Files
```

```
## Hubert J. Farnsworth (17.01.2045)
```

```
mv z_tmp/JL-2-76_S*_L001_R1_001.fastq.gz a_data/gz/JL-2-76_R1.fq.gz
mv z_tmp/JL-2-76_S*_L001_R2_001.fastq.gz a_data/gz/JL-2-76_R2.fq.gz
mv z_tmp/JL-2-86_S*_L001_R1_001.fastq.gz a_data/gz/JL-2-86_R1.fq.gz
mv z_tmp/JL-2-86_S*_L001_R2_001.fastq.gz a_data/gz/JL-2-86_R2.fq.gz
mv z_tmp/JL-2-96_S*_L001_R1_001.fastq.gz a_data/gz/JL-2-96_R1.fq.gz
mv z_tmp/JL-2-96_S*_L001_R2_001.fastq.gz a_data/gz/JL-2-96_R2.fq.gz
mv z_tmp/JL-3-53_S*_L001_R1_001.fastq.gz a_data/gz/JL-3-53_R1.fq.gz
mv z_tmp/JL-3-53_S*_L001_R2_001.fastq.gz a_data/gz/JL-3-53_R2.fq.gz
mv z_tmp/JL-3-63_S*_L001_R1_001.fastq.gz a_data/gz/JL-3-63_R1.fq.gz
mv z_tmp/JL-3-63_S*_L001_R2_001.fastq.gz a_data/gz/JL-3-63_R2.fq.gz
mv z_tmp/JL-3-73_S*_L001_R1_001.fastq.gz a_data/gz/JL-3-73_R1.fq.gz
mv z_tmp/JL-3-73_S*_L001_R2_001.fastq.gz a_data/gz/JL-3-73_R2.fq.gz
mv z_tmp/JL-4-20_S*_L001_R1_001.fastq.gz a_data/gz/JL-4-20_R1.fq.gz
mv z_tmp/JL-4-20_S*_L001_R2_001.fastq.gz a_data/gz/JL-4-20_R2.fq.gz
mv z_tmp/JL-4-23_S*_L001_R1_001.fastq.gz a_data/gz/JL-4-23_R1.fq.gz
mv z_tmp/JL-4-23_S*_L001_R2_001.fastq.gz a_data/gz/JL-4-23_R2.fq.gz
mv z_tmp/JL-4-33_S*_L001_R1_001.fastq.gz a_data/gz/JL-4-33_R1.fq.gz
mv z_tmp/JL-4-33_S*_L001_R2_001.fastq.gz a_data/gz/JL-4-33_R2.fq.gz
mv z_tmp/JL-4-50_S*_L001_R1_001.fastq.gz a_data/gz/JL-4-50_R1.fq.gz
mv z_tmp/JL-4-50_S*_L001_R2_001.fastq.gz a_data/gz/JL-4-50_R2.fq.gz
```

```
2019-10-26 app[java.2]: 126.0.132.125
```

```
.* (.*)\[(.*)\]:.*
```

```
[12]\d{3}-[01]\d-[0-3]\d ([^ \[\]]*?)\[( [^\]]*?)\]:.*
```

speed

(~42x faster)

Limits of (simple) find and replace

find: the

Whole word only

O Captain! my Captain! rise up and hear **the** bells;
Rise up-for you **the** flag is flung-for you **the** bugle trills;
For you bouquets and ribbon'd wreaths-for you **the** shores a-crowding;
For you **they** call, **the** swaying mass, **the**ir eager faces turning;
Here Captain! dear **father**!
This arm beneath your head;
It is some dream that on **the** deck,
You've fallen cold and dead.

by Walt Whitman

find & replace

Sample		Sample
A1-P-S	→	A1_P_S
A2-P-S	→	A2_P_S
A3-P-S	→	A3_P_S
A1-F-S	→	A1_F_S
A2-F-S	→	A2_F_S
A3-F-S	→	A3_F_S
A1-F-W	→	A1_F_W
A2-F-W	→	A2_F_W
A3-F-W	→	A3_F_W
B1-P-S	→	B1_P_S
B2-P-S	→	B2_P_S
B3-P-S	→	B3_P_S
B1-F-S	→	B1_F_S
B2-F-S	→	B2_F_S
B3-F-S	→	B3_F_S

Bioinformatics - RegEx

The screenshot shows a Microsoft Excel spreadsheet with a 'Replace' dialog box open. The spreadsheet contains a table with the following data:

Sample	Date	Tmp	pH	Colector
A1-P-S	17.07.12	28	7.2	DG
A2-P-S	17.07.12	28	7.3	DG
A3-P-S	17.07.12	28	7.7	DG
A1-F-S	17.07.12	25	6.3	DG
A2-F-S	17.07.12	25	6.5	DG
A3-F-S	17.07.12	25	6.6	DG
A1-F-W	16.11.26	7	7.4	MC
A2-F-W	16.11.26	7	7.1	MC
A3-F-W	16.11.26	7	6.9	MC
B1-P-S	17.07.18	29	7.1	DG
B2-P-S	17.07.18	29	7.2	DG
B3-P-S	17.07.18	29	7.2	DG
B1-F-S	17.07.20	27	6.6	MC
B2-F-S	17.07.20	27	6.7	MC
B3-F-S	17.07.20	27	6.5	MC

The 'Replace' dialog box is configured as follows:

- Find what: -
- Within: Sheet
- Search: By Rows
- Match case:
- Find entire cells only:
- Replace with: -

Buttons: Replace, Replace All, Close, Find Next. A hand cursor is pointing at the 'Replace All' button.

find & replace

Sample		Sample
A1-P-S	→	A1_P_S
A2-P-S	→	A2_P_S
A3-P-S	→	A3_P_S
A1-F-S	→	A1_F_S
A2-F-S	→	A2_F_S
A3-F-S	→	A3_F_S
A1-F-W	→	A1_F_W
A2-F-W	→	A2_F_W
A3-F-W	→	A3_F_W
B1-P-S	→	B1_P_S
B2-P-S	→	B2_P_S
B3-P-S	→	B3_P_S
B1-F-S	→	B1_F_S
B2-F-S	→	B2_F_S
B3-F-S	→	B3_F_S

regular **?** expression

Sample		Sample
A1-P-S	→	Sample_PS_A-1
A2-P-S	→	Sample_PS_A-2
A3-P-S	→	Sample_PS_A-3
A1-F-S	→	Sample_FS_A-1
A2-F-S	→	Sample_FS_A-2
A3-F-S	→	Sample_FS_A-3
A1-F-W	→	Sample_FW_A-1
A2-F-W	→	Sample_FW_A-2
A3-F-W	→	Sample_FW_A-3
B1-P-S	→	Sample_PS_B-1
B2-P-S	→	Sample_PS_B-2
B3-P-S	→	Sample_PS_B-3
B1-F-S	→	Sample_FS_B-1
B2-F-S	→	Sample_FS_B-2
B3-F-S	→	Sample_FS_B-3

Bioinformatics - RegEx

Sample		Sample
A1-P-S	→	Sample_PS_A-1
A2-P-S	→	Sample_PS_A-2
A3-P-S	→	Sample_PS_A-3
A1-F-S	→	Sample_FS_A-1
A2-F-S	→	Sample_FS_A-2
A3-F-S	→	Sample_FS_A-3
A1-F-W	→	Sample_FW_A-1
A2-F-W	→	Sample_FW_A-2
A3-F-W	→	Sample_FW_A-3
B1-P-S	→	Sample_PS_B-1
B2-P-S	→	Sample_PS_B-2
B3-P-S	→	Sample_PS_B-3
B1-F-S	→	Sample_FS_B-1
B2-F-S	→	Sample_FS_B-2
B3-F-S	→	Sample_FS_B-3

find: `(\w)(\d)-(\w)-(\w)`

replace*1: `Sample_`$3`$4_`$1`-`$2``

replace*2: `Sample_`3`4_`1`-`2``

*1 example for e.g. Atom

*2 example for e.g. TextWrangler

A **regular expression** (**regex** or **regexp** for short) is a special **text string for describing a search pattern**. You can think of regular expressions as wildcards on steroids.

```
\b[A-Z0-9._%+-]+@[A-Z0-9.-]+\.[A-Z]{2,4}\b
```

A regular expression "**engine**" is a piece of software that can process regular expressions, trying to match the pattern to the given string. Usually, the engine is part of a larger application and you do not access the engine directly.



The screenshot shows a text editor window titled "Project" with a file explorer on the left and a text editor on the right. The file explorer shows a directory structure with folders "Data", "Examples", and "Exercises". The "Exercises" folder is expanded, showing several files, including "EG17_BI_E2_RegEx_new.txt" which is selected. The text editor displays the content of this file, which is a document with RegEx exercises. The document starts with a header section followed by an introduction and three sections of exercises. A search dialog box is open at the bottom of the editor, showing a search query and the result "\$1 \$2".

```
1 ### =====
2 ### Course : Evolutionary Genetics UniBas (LV 25600-01 / HS2016)
3 ### Topic  : Regular Expressions
4 ### Version: 0.2 Jean-Claude Walser
5 ### =====
6
7 Following an introduction into regular expression. The introduction is divided into three
8 sections:
9 .....(1) Explore Examples - More Explore
10 .....(2) Exercises - Find Solution(s)
11 .....(3) Your Turn - Be Creative
12
13 In the explore part you have all the pieces (input, search term, replacement, and output)
14 and you have to figure out how it works. See the examples first to get an idea.
15
16 For the exercises input(s) (e.g. what you have) and output(s) (e.g. what you need)
17 are there (separated by ---). Try to figure out the find and replacements terms.
18 There are exercises with multiple steps. You start with the first input on top.
19 Convert it into the text below and use it as input for the next step to convert it into
20 the text below. Continue until you have reached the last text block. Solution are further
```

No results found for 'w{5} w{4} - \[([d{2}])\] (\w+|\w+ |\w+ |\w+ |\w+ |\w+ |\w+ |(\w+ |\w+))\.mp3'

Finding with Options: Regex, Case Sensitive, Within Current Selection

w{4} - \[([d{2}])\] (\w+|\w+ |\w+ |\w+ |\w+ |\w+ |\w+ |(\w+ |\w+)) no results

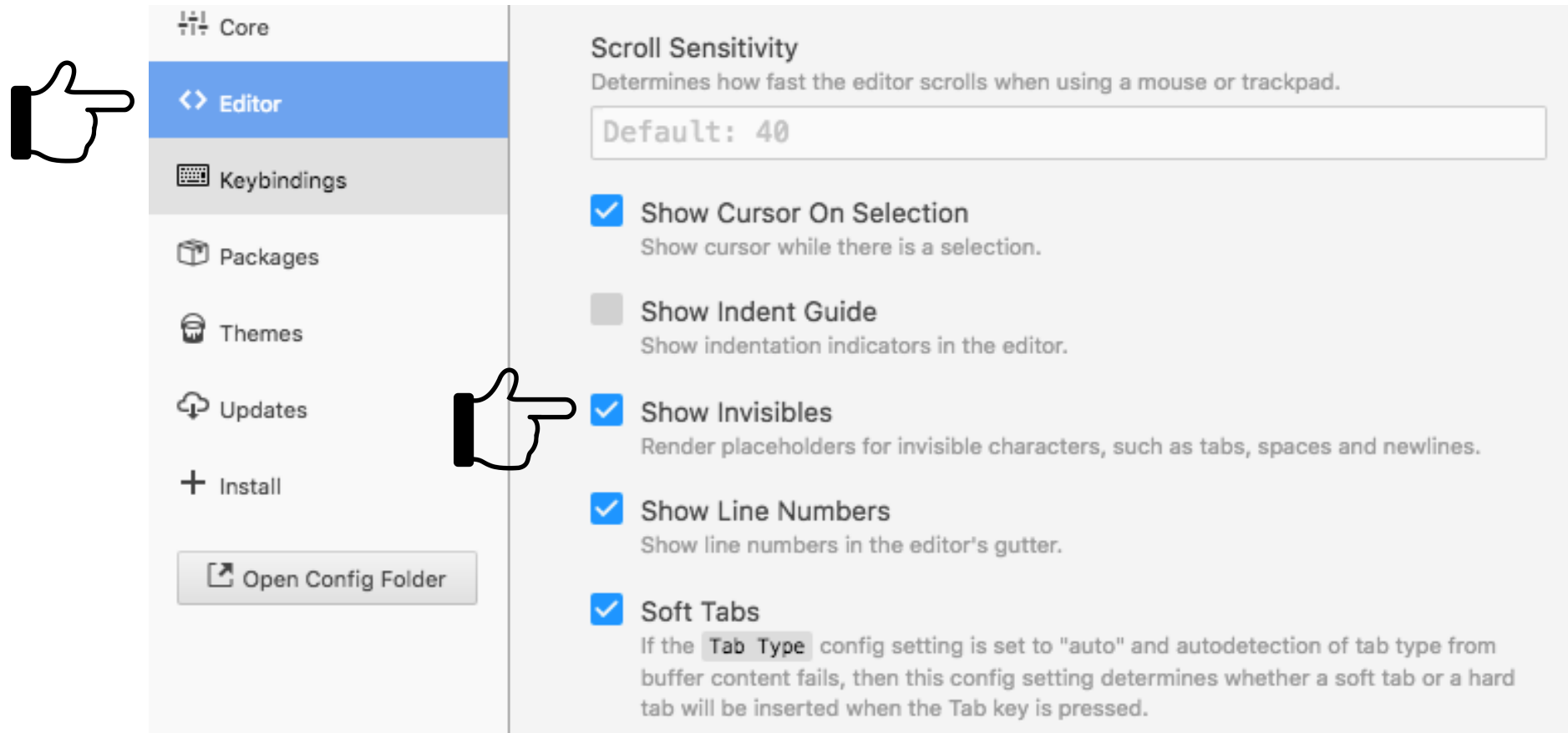
\$1 \$2

Replace Replace All

Exercises/EG17_BI_E2_RegEx_new.txt 1:1 LF UTF-8 Plain Text 0 files

<https://atom.io>

[Atom] > [Preferences...] > Settings



The screenshot shows the Atom editor's Settings window. The left sidebar contains the following categories: Core, Editor (selected), Keybindings, Packages, Themes, Updates, and Install. Below the sidebar is a button labeled 'Open Config Folder'. The main area displays the 'Scroll Sensitivity' setting, which is a text input field containing 'Default: 40'. Below this are several checked options:

- Show Cursor On Selection**
Show cursor while there is a selection.
- Show Indent Guide**
Show indentation indicators in the editor.
- Show Invisibles**
Render placeholders for invisible characters, such as tabs, spaces and newlines.
- Show Line Numbers**
Show line numbers in the editor's gutter.
- Soft Tabs**
If the `Tab Type` config setting is set to "auto" and autodetection of tab type from buffer content fails, then this config setting determines whether a soft tab or a hard tab will be inserted when the Tab key is pressed.

[Find] > [Find in Buffer] or [cmd] & [F]

Enable regex mode:



Bioinformatics - RegEx

regex	meaning
<code>\b</code>	word boundary

```
find: the\b
```

O Captain! my Captain! rise up and hear **the** bells;
Rise up-for you **the** flag is flung-for you **the** bugle trills;
For you bouquets and ribbon'd wreaths-for you **the** shores a-crowding;
For you they call, **the** swaying mass, their eager faces turning;
Here Captain! dear father!
This arm beneath your head;
It is some dream that on **the** deck,
You've fallen cold and dead.

by Walt Whitman

Bioinformatics - RegEx

regex	meaning
<code>\w</code>	word character (including letters, numbers, and underscore)
<code>\W</code>	not word character
<code>\w+</code>	one or more words
<code>\d</code>	digit
<code>\D</code>	not digit
<code>\d+</code>	one or more digits
<code>\b</code>	word boundary
<code>\t</code>	tab
<code>\s</code>	white space
<code>\r</code>	end-of-line (sometimes <code>\n</code>)

Bioinformatics - RegEx

regex	meaning
<code>^A</code>	begin with A
<code>A\$</code>	end with A
<code>[]</code>	character class/set
<code>[ACGT]</code>	set of character
<code>[A-Za-z]</code>	2 sets (ranges with metacharacter dash)
<code>[0-9\.]</code>	set of digits and decimal point
<code>[^TG]</code>	negation(s)
<code>{}</code>	number of matches
<code>{n}</code>	n matches
<code>{n,m}</code>	n-m matches
<code>{n,}</code>	n and more matches

Bioinformatics - RegEx

regex	meaning
()	search set
()	search set alternatives
\	escape character
\s	white space
\S	not white space
.	any character except new line
*	0 or more (e.g. .* whole line)
+	1 or more
?	0 or 1

match	skip
can	dan
man	ran
fan	pan

[cmf]an

[cmf].

^[cmf]

match	skip
hog	bog
dog	

[hd]og

^[hd]

match	skip
Ana	aax
Bob	bby
Cpc	zcz

[ABC]

[A-C]

^[A-C]

^[^a-z]

^[^abz]

match

wazzzzup

wazzzup

skip

wazup

waz{3}

waz{2,}

wazz. (!!!)

match
test.tmp
seq.tmp
help.tmp

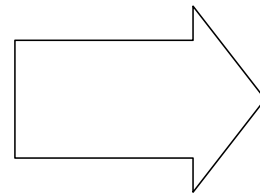
skip
seq.fa
tmp.tex
help.txt

*.tmp

tmp\$

(\w+).tmp\$

Mus musculus
Agalma elegans
Frillagalma vitiazi
Cordagalma tottoni
Shortia galacifolia



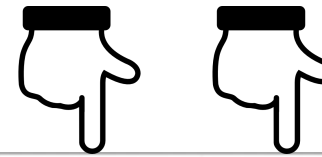
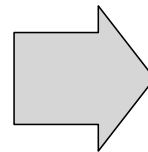
M. musculus
A. elegans
F. vitiazi
C. tottoni
S. galacifolia

Find: **(\w)\w+ (\w+)**

Replace: **\$1. \$2**

Bioinformatics - RegEx

A	B	C	D	E
Sample	Date	Tmp	pH	Colector
A1-P-S	17.07.12	28	7.2	DG
A2-P-S	17.07.12	28	7.3	DG
A3-P-S	17.07.12	28	7.7	DG
A1-F-S	17.07.12	25	6.3	DG
A2-F-S	17.07.12	25	6.5	DG
A3-F-S	17.07.12	25	6.6	DG
A1-F-W	16.11.26	7	7.4	MC
A2-F-W	16.11.26	7	7.1	MC
A3-F-W	16.11.26	7	6.9	MC
B1-P-S	17.07.18	29	7.1	DG
B2-P-S	17.07.18	29	7.2	DG
B3-P-S	17.07.18	29	7.2	DG
B1-F-S	17.07.20	27	6.6	MC
B2-F-S	17.07.20	27	6.7	MC
B3-F-S	17.07.20	27	6.5	MC

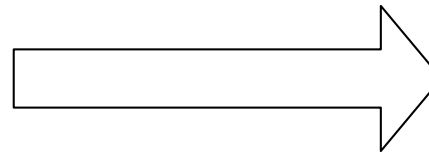


A	B	C	D	E
Sample	Date	Tmp	pH	Collector
Sample_PS_A-1	12.07.17	28	7.2	DG
Sample_PS_A-2	12.07.17	28	7.3	DG
Sample_PS_A-3	12.07.17	28	7.7	DG
Sample_FS_A-1	12.07.17	25	6.3	DG
Sample_FS_A-2	12.07.17	25	6.5	DG
Sample_FS_A-3	12.07.17	25	6.6	DG
Sample_FW_A-1	26.11.16	7	7.4	MC
Sample_FW_A-2	26.11.16	7	7.1	MC
Sample_FW_A-3	26.11.16	7	6.9	MC
Sample_PS_B-1	18.07.17	29	7.1	DG
Sample_PS_B-2	18.07.17	29	7.2	DG
Sample_PS_B-3	18.07.17	29	7.2	DG
Sample_FS_B-1	20.07.17	27	6.6	MC
Sample_FS_B-2	20.07.17	27	6.7	MC
Sample_FS_B-3	20.07.17	27	6.5	MC

Example_Table.xls

A	B	C	D	E
Sample	Date	Tmp	pH	Colector
A1-P-S	17.07.12	28	7.2	DG
A2-P-S	17.07.12	28	7.3	DG
A3-P-S	17.07.12	28	7.7	DG
A1-F-S	17.07.12	25	6.3	DG
A2-F-S	17.07.12	25	6.5	DG
A3-F-S	17.07.12	25	6.6	DG
A1-F-W	16.11.26	7	7.4	MC
A2-F-W	16.11.26	7	7.1	MC
A3-F-W	16.11.26	7	6.9	MC
B1-P-S	17.07.18	29	7.1	DG
B2-P-S	17.07.18	29	7.2	DG
B3-P-S	17.07.18	29	7.2	DG
B1-F-S	17.07.20	27	6.6	MC
B2-F-S	17.07.20	27	6.7	MC
B3-F-S	17.07.20	27	6.5	MC

Save As...
Tab Delimited Text (.txt)

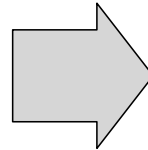


Example_Table.txt

```
Sample» Date» Tmp»pH» Colector»  
A1-P-S» 17.07.12» 28» 7.2»DG»  
A2-P-S» 17.07.12» 28» 7.3»DG»  
A3-P-S» 17.07.12» 28» 7.7»DG»  
A1-F-S» 17.07.12» 25» 6.3»DG»  
A2-F-S» 17.07.12» 25» 6.5»DG»  
A3-F-S» 17.07.12» 25» 6.6»DG»  
A1-F-W» 16.11.26» 7»7.4»MC»  
A2-F-W» 16.11.26» 7»7.1»MC»  
A3-F-W» 16.11.26» 7»6.9»MC»  
B1-P-S» 17.07.18» 29» 7.1»DG»  
B2-P-S» 17.07.18» 29» 7.2»DG»  
B3-P-S» 17.07.18» 29» 7.2»DG»  
B1-F-S» 17.07.20» 27» 6.6»MC»  
B2-F-S» 17.07.20» 27» 6.7»MC»  
B3-F-S» 17.07.20» 27» 6.5»MC»
```

Bioinformatics - RegEx

```
Sample» Date» Tmp» pH» Collector»  
A1-P-S» 17.07.12» 28» 7.2» DG»  
A2-P-S» 17.07.12» 28» 7.3» DG»  
A3-P-S» 17.07.12» 28» 7.7» DG»  
A1-F-S» 17.07.12» 25» 6.3» DG»  
A2-F-S» 17.07.12» 25» 6.5» DG»  
A3-F-S» 17.07.12» 25» 6.6» DG»  
A1-F-W» 16.11.26» 7» 7.4» MC»  
A2-F-W» 16.11.26» 7» 7.1» MC»  
A3-F-W» 16.11.26» 7» 6.9» MC»  
B1-P-S» 17.07.18» 29» 7.1» DG»  
B2-P-S» 17.07.18» 29» 7.2» DG»  
B3-P-S» 17.07.18» 29» 7.2» DG»  
B1-F-S» 17.07.20» 27» 6.6» MC»  
B2-F-S» 17.07.20» 27» 6.7» MC»  
B3-F-S» 17.07.20» 27» 6.5» MC»
```



```
Sample» Date» Tmp» pH» Collector»  
Sample_PS_A-1» 12.07.17» 28» 7.2» DG»  
Sample_PS_A-2» 12.07.17» 28» 7.3» DG»  
Sample_PS_A-3» 12.07.17» 28» 7.7» DG»  
Sample_FS_A-1» 12.07.17» 25» 6.3» DG»  
Sample_FS_A-2» 12.07.17» 25» 6.5» DG»  
Sample_FS_A-3» 12.07.17» 25» 6.6» DG»  
Sample_FW_A-1» 26.11.16» 7» 7.4» MC»  
Sample_FW_A-2» 26.11.16» 7» 7.1» MC»  
Sample_FW_A-3» 26.11.16» 7» 6.9» MC»  
Sample_PS_B-1» 18.07.17» 29» 7.1» DG»  
Sample_PS_B-2» 18.07.17» 29» 7.2» DG»  
Sample_PS_B-3» 18.07.17» 29» 7.2» DG»  
Sample_FS_B-1» 20.07.17» 27» 6.6» MC»  
Sample_FS_B-2» 20.07.17» 27» 6.7» MC»  
Sample_FS_B-3» 20.07.17» 27» 6.5» MC»
```

find: **(\w)(\d)-(\w)-(\w)\t(\d{2}).(\d{2}).(\d+)\t(\d+)\t(\d.\d)\t(\w+)**

replace: **Sample_** $\$3\4 **_** $\$1$ **-** $\$2$ **\t** $\$7$ **.** $\$6$ **.** $\$5$ **\t** $\$8$ **\t** $\$9$ **\t** $\$10$