

Evolutionary Genetics

LV 25600-01 | Lecture with exercises | 4KP

Regular Expressions

Jean-Claude Walser

jean-claude.walser [at] env.ethz.ch

RegEx

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

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.

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 table of sample data. A 'Replace' dialog box is open, allowing for a search and replace operation across the entire sheet. The search criteria are set to find hyphens ('-') within the sheet, searching by rows. The 'Replace All' button is highlighted by a hand cursor, indicating the next action to be taken.

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

Replace Dialog Box Settings:

- Find what: -
- Within: Sheet
- Search: By Rows
- Match case: ☐
- Find entire cells only: ☐
- Replace with: -
- Buttons: Replace, Replace All, Close, Find Next

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

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

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, **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

regex	meaning
\b	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

Sample	Read	Format
JL-2-76_S7_L001_R2_001.fastq.gz		
ID	Lane	Cell
Archive		

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

```
#!/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
```

JL-2-76_S7_L001_R2_001.fastq.gz

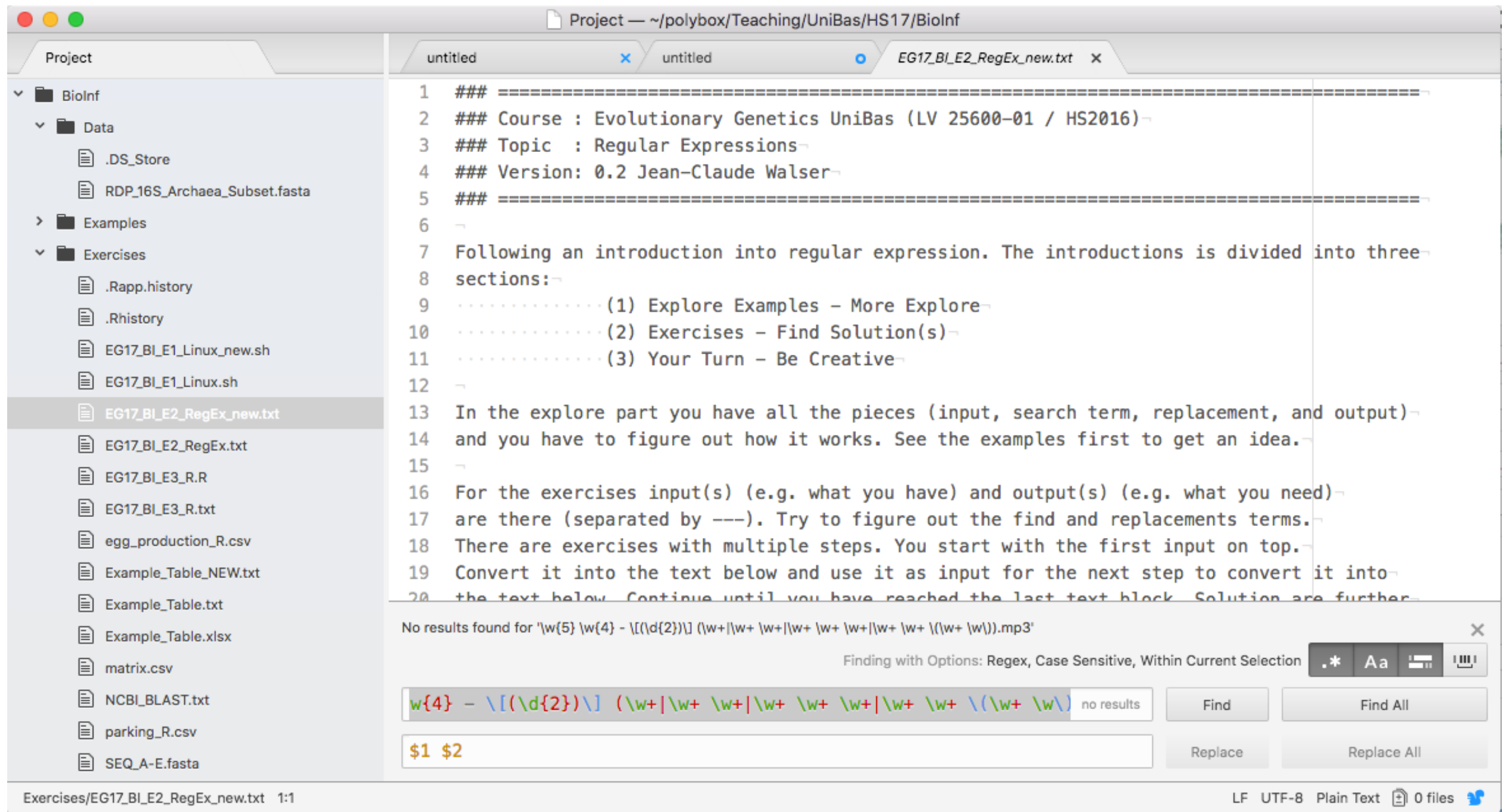
Find

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

Replace

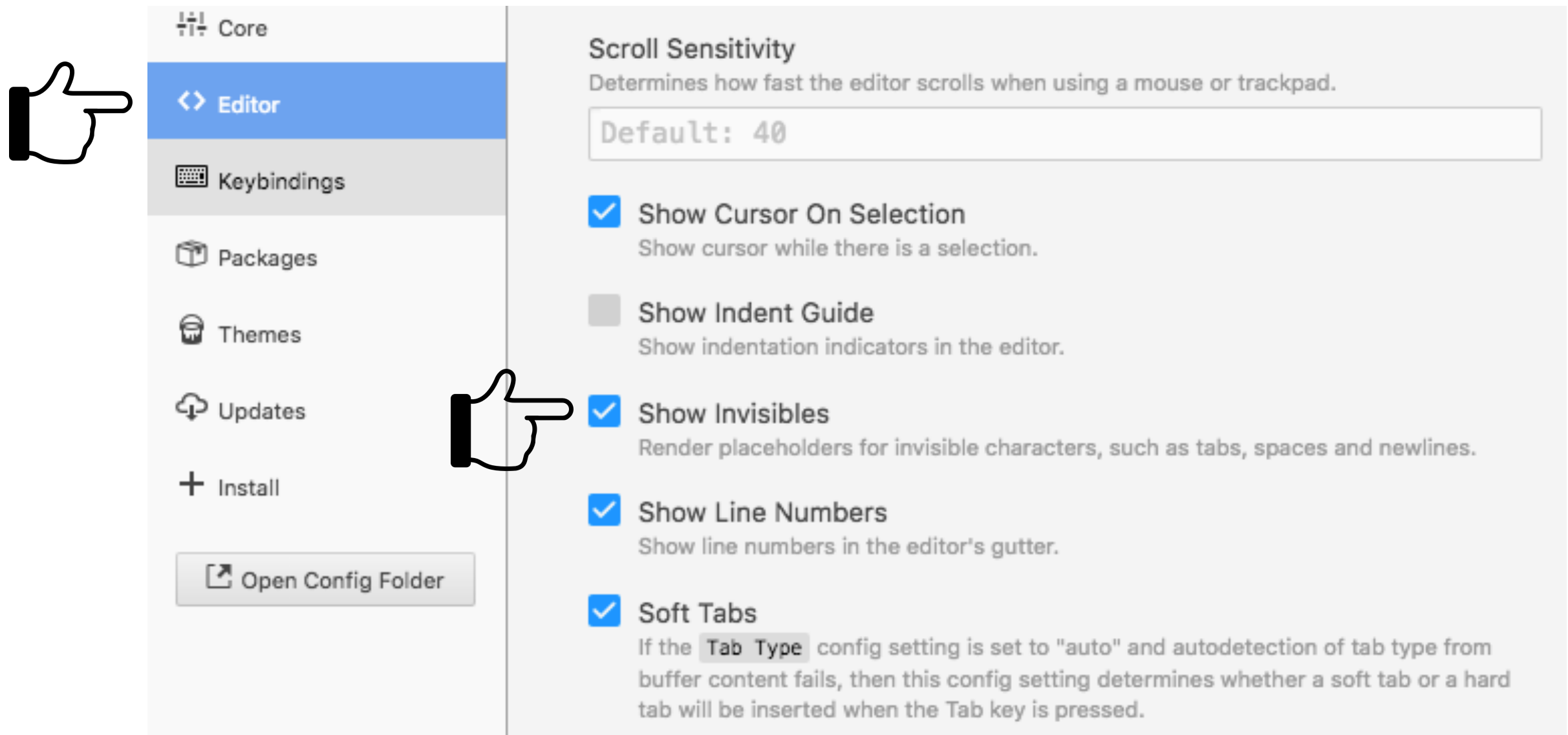
`a_data/gz/\b1\b2.fq.gz`

JL-2-76_R2.fq.gz



<https://atom.io>

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



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

Enable regex mode:

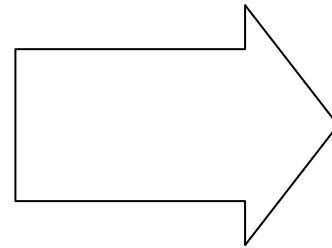


regex	meaning
\w	word character (including letters, numbers, and underscore)
\W	not word character
\w+	one or more words
\d	digit
\D	not digit
\d+	one or more digits
\b	word boundary
\t	tab
\s	white space
\r	end-of-line (sometimes \n)

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

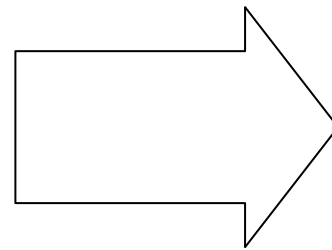
regex	meaning
()	search set
()	search set alternatives
\	escape character
\s	white space
\S	not white space
.	any character except new line
*	0 or more (e.g. .* whole line)
+	1 or more
?	0 or 1

Mus musculus
Agalma elegans
Frillagalma vitiazi
Cordagalma tottoni
Shortia galacifolia



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

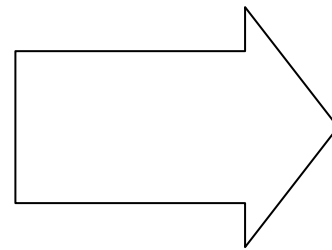
Mus musculus
Agalma elegans
Frillagalma vitiazi
Cordagalma tottoni
Shortia galacifolia



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

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

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**



match

skip

can

dan

man

ran

fan

pan

match

can

man

fan

skip

dan

ran

pan

[cmf]an

[cmf].

^[cmf]



match	skip
Ana	aax
Bob	bby
Cpc	zcZ

match	skip
Ana	aax
Bob	bby
Cpc	zcz

[ABC]

[A-C]

^[A-C]

^[^a-z]

^[^abz]



match

wazzzzup

wazzzup

skip

wazup

match

wazzzzup

wazzzup

skip

wazup

waz{3}

waz{2,}

wazz. (!!!)



match

test.tmp

seq.tmp

help.tmp

skip

seq.fa

tmp.tex

help.txt

match

test.tmp

seq.tmp

help.tmp

skip

seq.fa

tmp.tex

help.txt

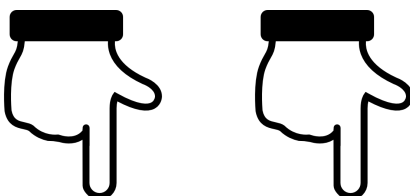
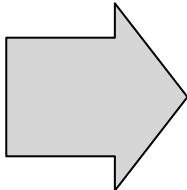
`(\w+).tmp$`

`*.tmp`

`tmp$`

`.+\.tmp`

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

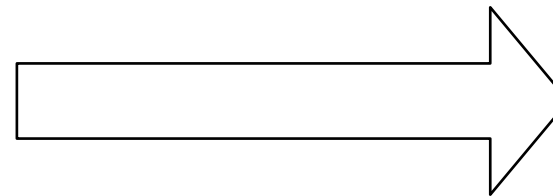


A	B	C	D	E
Sample	Date	Tmp	pH	Collector
Sample_PS_A-1	12.07.17	28	7.2	DG
Sample_PS_A-2	12.07.17	28	7.3	DG
Sample_PS_A-3	12.07.17	28	7.7	DG
Sample_FS_A-1	12.07.17	25	6.3	DG
Sample_FS_A-2	12.07.17	25	6.5	DG
Sample_FS_A-3	12.07.17	25	6.6	DG
Sample_FW_A-1	26.11.16	7	7.4	MC
Sample_FW_A-2	26.11.16	7	7.1	MC
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)

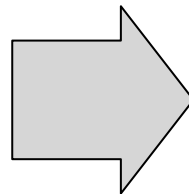


Example_Table.txt

```
Sample» Date» Tmp» pH» 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
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**

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

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

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

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

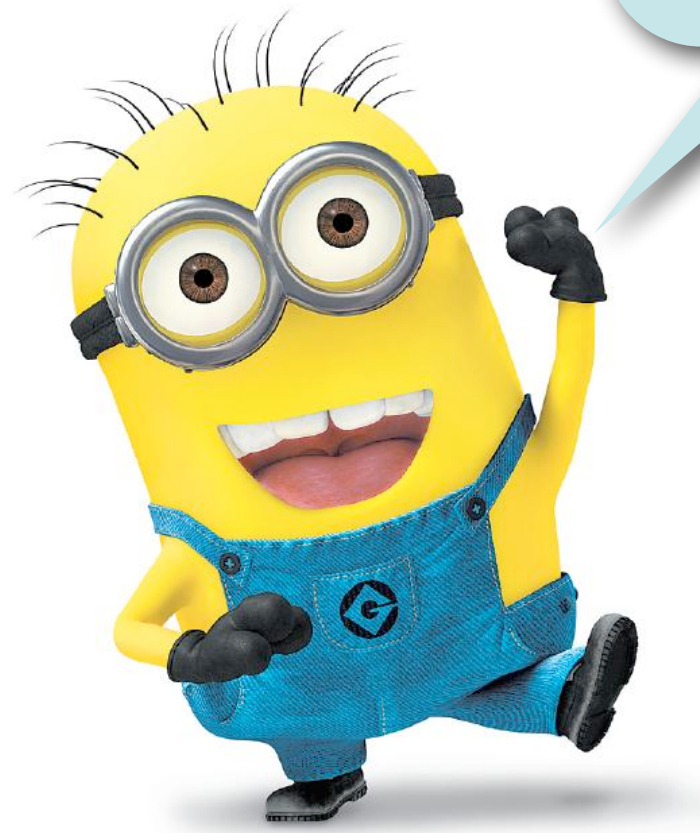
```
.* (.*)\[(.*)\]:.*
```

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

speed

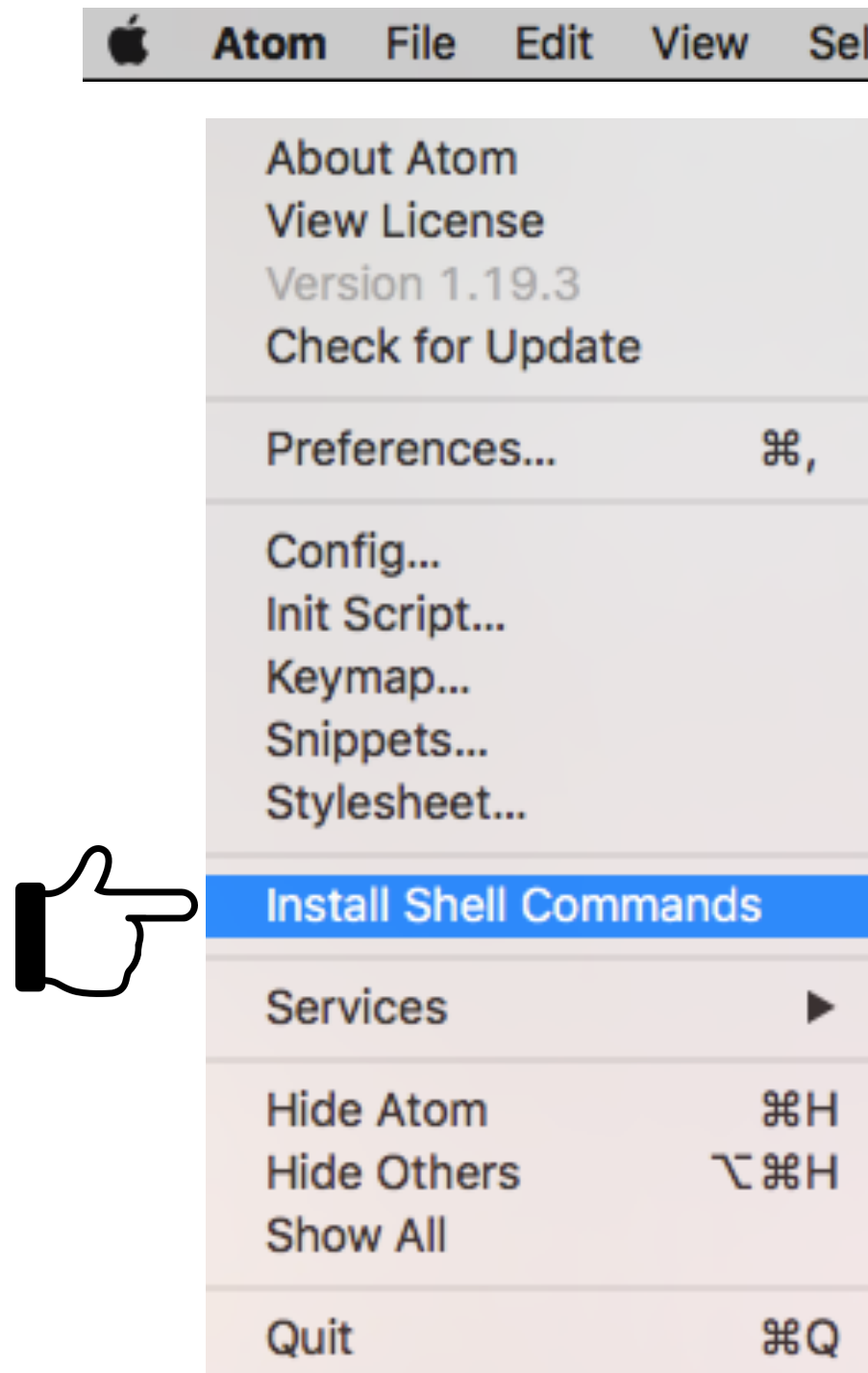
(~42x faster)





Your turn!

**MORE
THINGS
CONSIDERED**



```
# Terminal command  
atom EG17_BI_E2_RegEx.txt  
# or  
atom ./Exercices  
# will open new window and  
# list all files inside
```