Computer classes should teach regular expressions to kids

Cory Doctorow at 10:39 am Tue, Dec 4

regular expressions

A pattern that strings can match

Like using “find”

regular expressions

A pattern that strings can match

Like using “find”

regular expressions

A pattern that strings can match

Warning: notation can get ugly

regular expressions

A pattern that strings can match

Warning: notation can get ugly

Writing patterns for strings *as* strings...

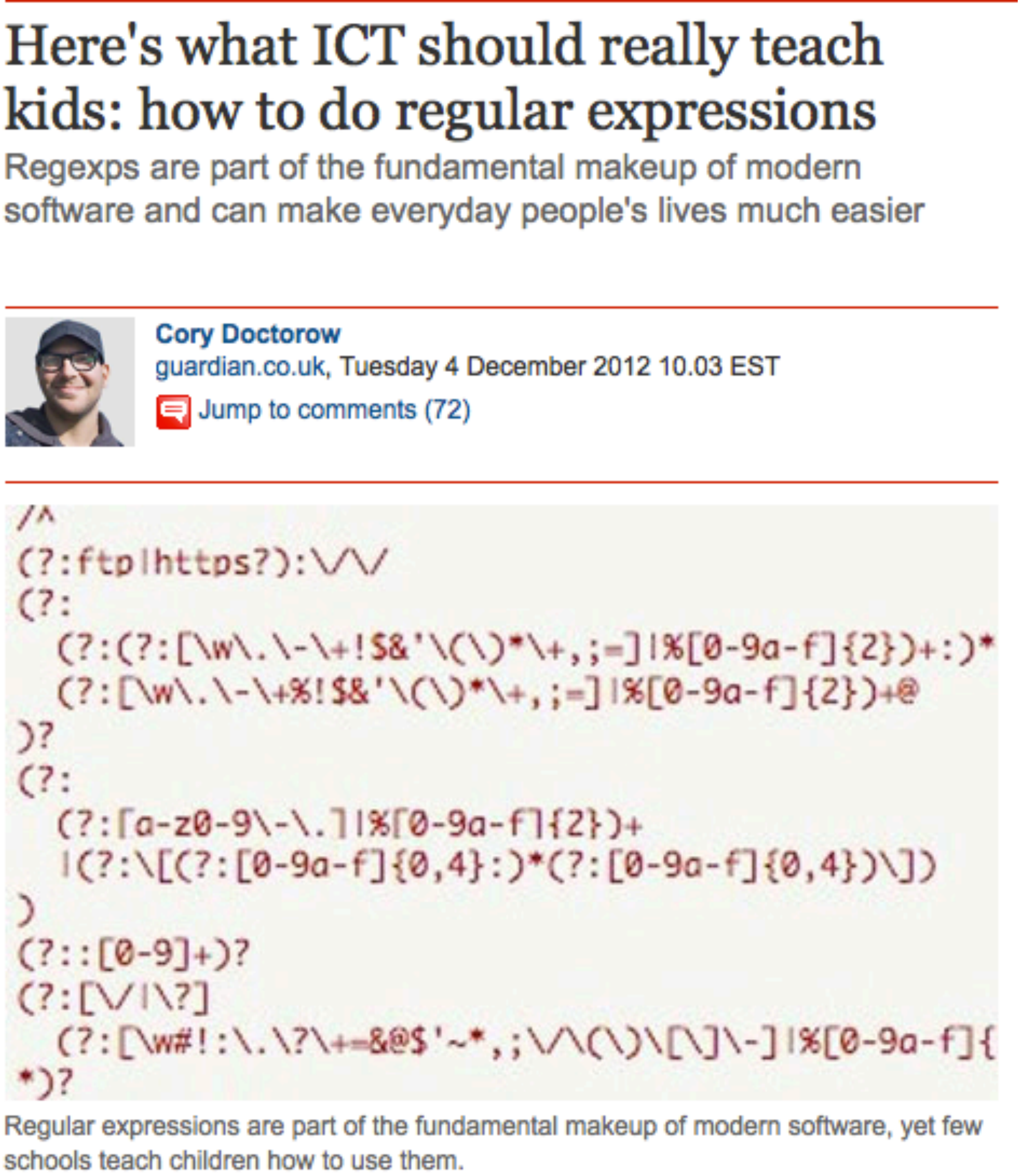
regular expressions

A pattern that strings can match

Warning: notation can get ugly

Writing patterns for strings *as* strings…

...using only the symbols on the keyboard (instead of inventing new symbols like mathematicians do)



Here's what ICT should really teach kids: how to do regular expressions

Regexps are part of the fundamental makeup of modern software and can make everyday people's lives much easier

Cory Doctorow
guardian.co.uk, Tuesday 4 December 2012 10.03 EST
Jump to comments (72)

```
/^
(?:ftp|https?):\/\/
(?:
  (?:(?:[\w\.\-\+!$&'\(\)*\+,;=]|%[0-9a-f]{2})+:)*
  (?:[\w\.\-\+%!$&'\(\)*\+,;=]|%[0-9a-f]{2})+@
)?
(?:
  (?:[a-z0-9\-\.]|%[0-9a-f]{2})+
  |(?:\[(?:[0-9a-f]{0,4}:)*(?:[0-9a-f]{0,4})\])
)
(?::[0-9]+)?
(?:[\/|\?]
  (?:[\w#!:\.\?\+=&@$'~*,;\/\(\)\[\]\-]|%[0-9a-f]{
*)?
```

Regular expressions are part of the fundamental makeup of modern software, yet few schools teach children how to use them.

Why not just use “find”?

in python:

```
searchstring=“thing I’m lookin for”
for lin in file:
    if searchstring in lin:
        print lin
```

Why not just use “find”?

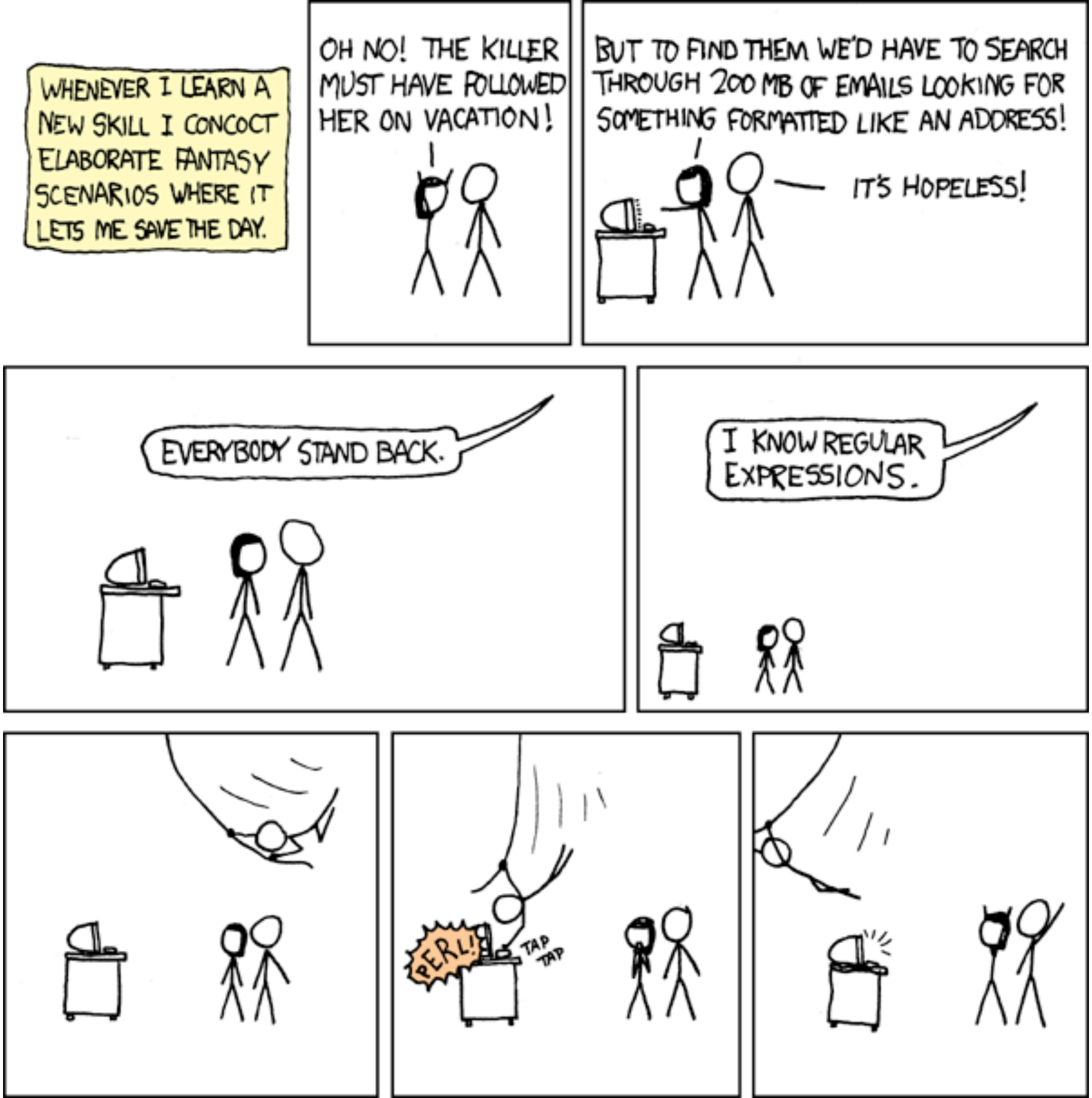
in python:

```
searchstring=“thing I’m lookin for”
for lin in file:
    if searchstring in lin:
        print lin
```

FLEXIBILITY

Regular Expressions are useful if you only *kind of* know what you are looking for.

especially if you know how it would be formatted


WHENEVER I LEARN A NEW SKILL I CONCOCT ELABORATE FANTASY SCENARIOS WHERE IT LETS ME SAVE THE DAY.
OH NO! THE KILLER MUST HAVE FOLLOWED HER ON VACATION!
BUT TO FIND THEM WE'D HAVE TO SEARCH THROUGH 200 MB OF EMAILS LOOKING FOR SOMETHING FORMATTED LIKE AN ADDRESS!
IT'S HOPELESS!
EVERYBODY STAND BACK.
I KNOW REGULAR EXPRESSIONS.
PERL!
TAP TAP

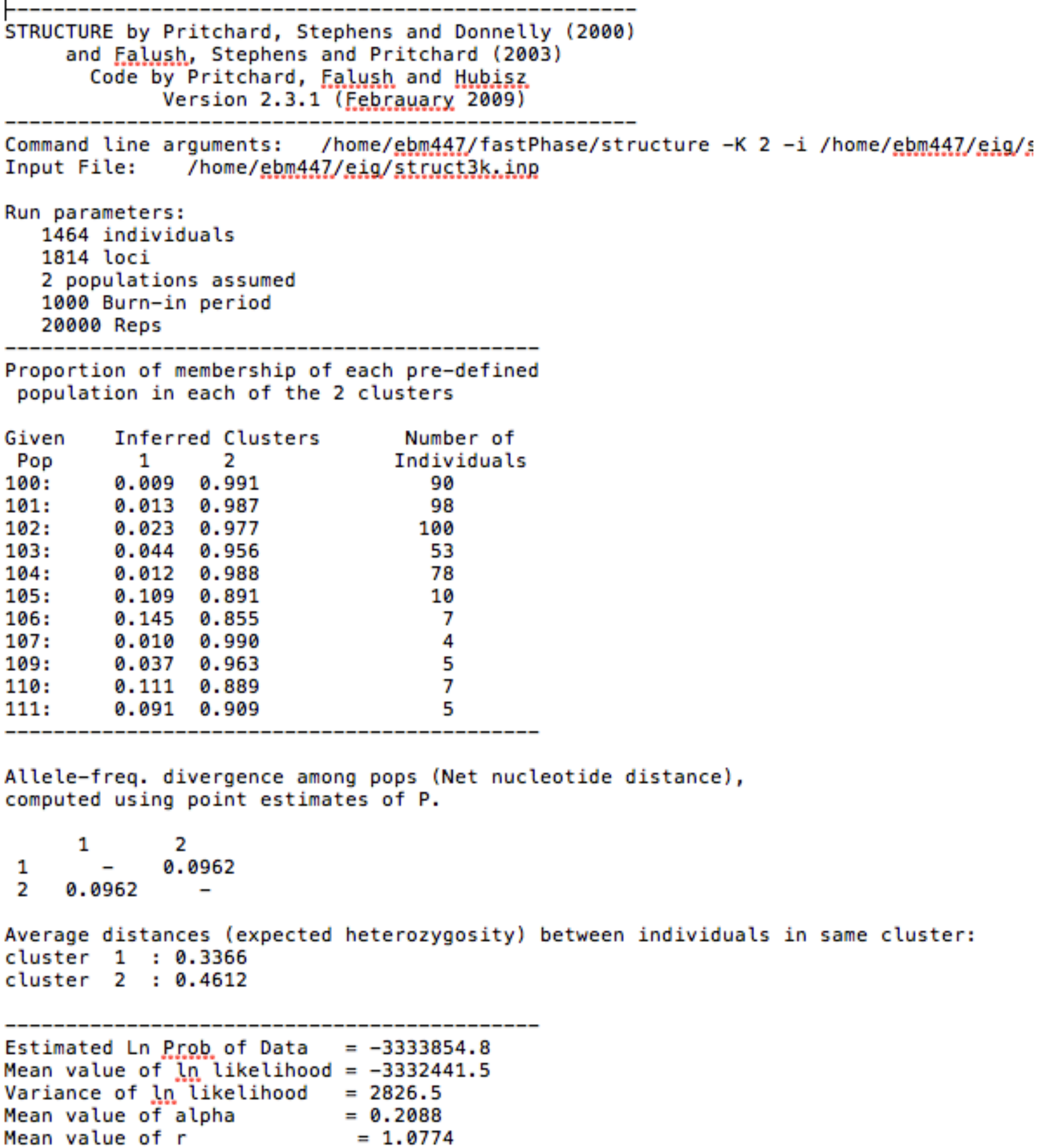
More common than you might think!

e.g.

pull filenames out of textfiles

find all the phone numbers in an email

search for a range of dates



```
----------------------------------------------------
STRUCTURE by Pritchard, Stephens and Donnelly (2000)
     and Falush, Stephens and Pritchard (2003)
       Code by Pritchard, Falush and Hubisz
             Version 2.3.1 (Febrauary 2009)
----------------------------------------------------
Command line arguments:   /home/ebm447/fastPhase/structure -K 2 -i /home/ebm447/eig/s
Input File:    /home/ebm447/eig/struct3k.inp

Run parameters:
   1464 individuals
   1814 loci
   2 populations assumed
   1000 Burn-in period
   20000 Reps
--------------------------------------------
Proportion of membership of each pre-defined
 population in each of the 2 clusters

Given    Inferred Clusters       Number of
 Pop       1      2             Individuals
100:     0.009  0.991              90
101:     0.013  0.987              98
102:     0.023  0.977             100
103:     0.044  0.956              53
104:     0.012  0.988              78
105:     0.109  0.891              10
106:     0.145  0.855               7
107:     0.010  0.990               4
109:     0.037  0.963               5
110:     0.111  0.889               7
111:     0.091  0.909               5
--------------------------------------------

Allele-freq. divergence among pops (Net nucleotide distance),
computed using point estimates of P.

      1       2
 1      -    0.0962
 2   0.0962     -

Average distances (expected heterozygosity) between individuals in same cluster:
cluster  1  : 0.3366
cluster  2  : 0.4612

--------------------------------------------
Estimated Ln Prob of Data   = -3333854.8
Mean value of ln likelihood = -3332441.5
Variance of ln likelihood   = 2826.5
Mean value of alpha         = 0.2088
Mean value of r              = 1.0774
```

download structureoutput.txt

(from the webpage)

RegExPal demo
regexpal.com
structureoutput.txt

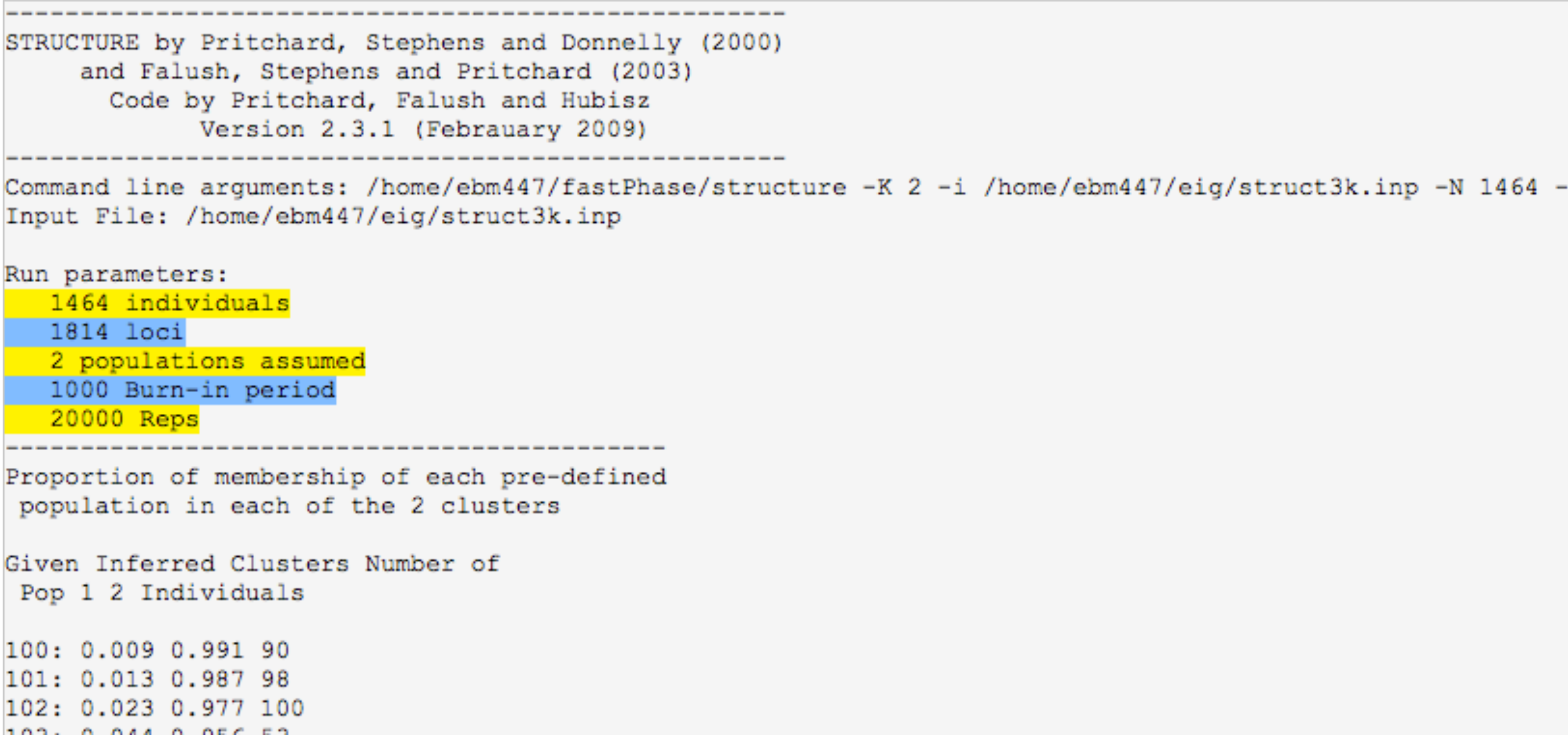
Cheat sheet from regexpal

Regular characters act normally

.	Any character except newline.
\.	A period (and so on for *, \(, \\, etc.)
^	The start of the string.
$	The end of the string.
\d,\w,\s	A digit, word character [A-Za-z0-9_], or whitespace.
\D,\W,\S	Anything except a digit, word character, or whitespace.
[abc]	Character a, b, or c. [a-z] a through z.
[^abc]	Any character except a, b, or c.
aa|bb	Either aa or bb.
?	Zero or one of the preceding element.
*	Zero or more of the preceding element.
+	One or more of the preceding element.
{*n*}	Exactly *n* of the preceding element.
{*n*,}	*n* or more of the preceding element.
{*m*,*n*}	Between *m* and *n* of the preceding element. ??,*?,+?,

Mini-exercise!

You want to search through a bunch of different output files and get the run parameters for each one.
Write a regular expression in regexpal that will find only the five lines following “Run Parameters”



```
----------------------------------------------------
STRUCTURE by Pritchard, Stephens and Donnelly (2000)
     and Falush, Stephens and Pritchard (2003)
       Code by Pritchard, Falush and Hubisz
             Version 2.3.1 (Febrauary 2009)
----------------------------------------------------
Command line arguments: /home/ebm447/fastPhase/structure -K 2 -i /home/ebm447/eig/struct3k.inp -N 1464 -
Input File: /home/ebm447/eig/struct3k.inp

Run parameters:
   1464 individuals
   1814 loci
   2 populations assumed
   1000 Burn-in period
   20000 Reps
--------------------------------------------
Proportion of membership of each pre-defined
 population in each of the 2 clusters

Given Inferred Clusters Number of
 Pop 1 2 Individuals

100: 0.009 0.991 90
101: 0.013 0.987 98
102: 0.023 0.977 100
103: 0.044 0.956 53
```

My answer:

```
^\s+\d+ \w+.+
```

ipython notebook
reg ex demo using structureoutput.txt