

Week 8 assignment is, inevitably, about Regular Expressions and OpenRefine
Upload a text file or a PDF with your answers/solutions to the problems below. Beware of making the submission legible and understandable to another reader; for example, consider using the "Save regex" functionality in regex101.com, which allows you to create a link out of your solution and share the link for easy use by your colleagues. Remember that you can elaborate solutions in groups, but need to submit individually.

1. What regular expressions do you use to extract all the dates in this blurb:

<http://bit.ly/regexexercise2> and to put them into the following format YYYY-MM-DD ?
[regex101: build, test, and debug regex](#)

REGULAR EXPRESSION 6 matches (42 steps, 115µs)

`/ \d{1,2}.\d{1,2}.\s?\d+ /gm`

TEST STRING

Juan Ponce de León sights Florida for the first time, on 3.27, 1513^d
Giovanni da Verrazzano explored the Atlantic coast of North America under French employ, on 4.17.1524^d
The Roanoke Colony was found deserted, on 8/15/1590^d
John Smith founded the Jamestown settlement, on 5/14, 1607^d
The Dutch laid claim to the territories of New Netherland, on 11.11.1614^d
The Massachusetts Bay Colony founded, on 3-4-1629^d

REGULAR EXPRESSION 6 matches (78 steps, 140µs)

`/ (\d{1,2}).(\d{1,2}).(\s?\d+) /gm`

TEST STRING

Juan Ponce de León sights Florida for the first time, on 3.27, 1513^d
Giovanni da Verrazzano explored the Atlantic coast of North America under French employ, on 4.17.1524^d
The Roanoke Colony was found deserted, on 8/15/1590^d
John Smith founded the Jamestown settlement, on 5/14, 1607^d
The Dutch laid claim to the territories of New Netherland, on 11.11.1614^d
The Massachusetts Bay Colony founded, on 3-4-1629^d

7:1

SUBSTITUTION success (315µs)

`$3-$1-$2`

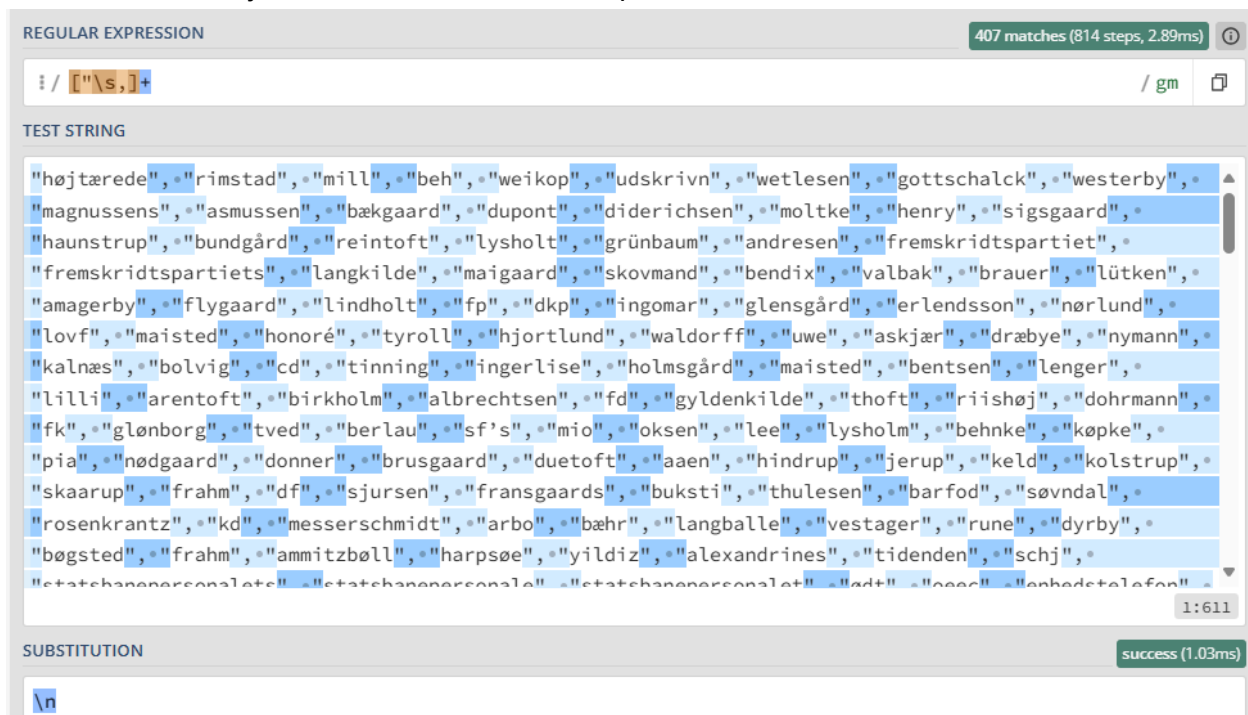
- ```
1 / (.+)(
2)
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60
61
62
63
64
65
66
67
68
69
70
71
72
73
74
75
76
77
78
79
80
81
82
83
84
85
86
87
88
89
90
91
92
93
94
95
96
97
98
99
100
101
102
103
104
105
106
107
108
109
110
111
112
113
114
115
116
117
118
119
120
121
122
123
124
125
126
127
128
129
130
131
132
133
134
135
136
137
138
139
140
141
142
143
144
145
146
147
148
149
150
151
152
153
154
155
156
157
158
159
160
161
162
163
164
165
166
167
168
169
170
171
172
173
174
175
176
177
178
179
180
181
182
183
184
185
186
187
188
189
190
191
192
193
194
195
196
197
198
199
200
201
202
203
204
205
206
207
208
209
210
211
212
213
214
215
216
217
218
219
220
221
222
223
224
225
226
227
228
229
230
231
232
233
234
235
236
237
238
239
240
241
242
243
244
245
246
247
248
249
250
251
252
253
254
255
256
257
258
259
260
261
262
263
264
265
266
267
268
269
270
271
272
273
274
275
276
277
278
279
280
281
282
283
284
285
286
287
288
289
290
291
292
293
294
295
296
297
298
299
300
301
302
303
304
305
306
307
308
309
310
311
312
313
314
315
316
317
318
319
320
321
322
323
324
325
326
327
328
329
330
331
332
333
334
335
336
337
338
339
340
341
342
343
344
345
346
347
348
349
350
351
352
353
354
355
356
357
358
359
360
361
362
363
364
365
366
367
368
369
370
371
372
373
374
375
376
377
378
379
380
381
382
383
384
385
386
387
388
389
390
391
392
393
394
395
396
397
398
399
400
401
402
403
404
405
406
407
408
409
410
411
412
413
414
415
416
417
418
419
420
421
422
423
424
425
426
427
428
429
430
431
432
433
434
435
436
437
438
439
440
441
442
443
444
445
446
447
448
449
450
451
452
453
454
455
456
457
458
459
460
461
462
463
464
465
466
467
468
469
470
471
472
473
474
475
476
477
478
479
480
481
482
483
484
485
486
487
488
489
490
491
492
493
494
495
496
497
498
499
500
501
502
503
504
505
506
507
508
509
510
511
512
513
514
515
516
517
518
519
520
521
522
523
524
525
526
527
528
529
530
531
532
533
534
535
536
537
538
539
540
541
542
543
544
545
546
547
548
549
550
551
552
553
554
555
556
557
558
559
560
561
562
563
564
565
566
567
568
569
570
571
572
573
574
575
576
577
578
579
580
581
582
583
584
585
586
587
588
589
590
591
592
593
594
595
596
597
598
599
600
601
602
603
604
605
606
607
608
609
610
611
612
613
614
615
616
617
618
619
620
621
622
623
624
625
626
627
628
629
630
631
632
633
634
635
636
637
638
639
640
641
642
643
644
645
646
647
648
649
650
651
652
653
654
655
656
657
658
659
660
661
662
663
664
665
666
667
668
669
670
671
672
673
674
675
676
677
678
679
680
681
682
683
684
685
686
687
688
689
690
691
692
693
694
695
696
697
698
699
700
701
702
703
704
705
706
707
708
709
710
711
712
713
714
715
716
717
718
719
720
721
722
723
724
725
726
727
728
729
730
731
732
733
734
735
736
737
738
739
740
741
742
743
744
745
746
747
748
749
750
751
752
753
754
755
756
757
758
759
760
761
762
763
764
765
766
767
768
769
770
771
772
773
774
775
776
777
778
779
780
781
782
783
784
785
786
787
788
789
790
791
792
793
794
795
796
797
798
799
800
801
802
803
804
805
806
807
808
809
810
811
812
813
814
815
816
817
818
819
820
821
822
823
824
825
826
827
828
829
830
831
832
833

```

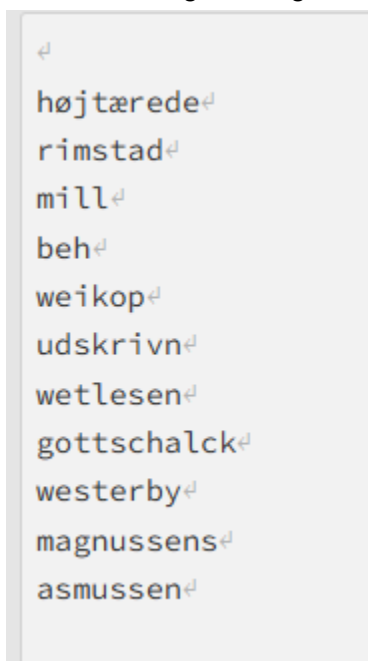
2. Then take the stopwordlist from R <http://bit.ly/regexexercise4> and convert it into a Voyant list (words on separate line without interpunction)<sup>4</sup>

Herefter skal vi tage stop ord listen, og omstrukturere daterne, så vi kan bruge den i voyant. Her kopiere man ordene og sætter dem ind i regex:

Derefter skal man fjerne alle værdierne, som separere ordene:



Denne kodning skulle gerne resultere i at ordene bliver som følgende:



<https://regex101.com/r/vGNdvw/1>

Herefter skal man kopiere ordene Ctrl-A + Ctrl-C

Så skal man kopiere ordene i en txt fil, og så skulle den gerne være klar til at bruge i voyant

3. Does OpenRefine alter the raw data during sorting and filtering?

Nej, openrefine kan komprimere data til en mere overskuelig sammensætning. Men den laver ikke om på dataene.

4. Fix the [interviews dataset](#) in OpenRefine enough to answer this question: "Which two months are reported as the most water-deprived/dryest by the interviewed farmer households?"

- Først har vi gjort cellerne under "months\_no\_water" mere overskuelige ved at fjerne "[ ]" og de mellemrum der var. Det har vi gjort ved at trykke "Edit cell" → "Transform" → Skrive: `value.replace("[", "").replace("]", "").replace(" ", "")`
- Da flere af cellerne indeholder flere datasvar har vi splittet hvert datasvar og derefter grupperet dem, dog uden at selve dataen i cellerne ændrer sig. Det har vi gjort ved at trykke "Facet" → "Custom text facet" → Skrive: `value.split(",")`
- Vi kan herefter så aflæse hvilke to måneder i datasættet der var tørrest ved at aflæse hvilke to datasvar der er rapporteret flest gange
- Ud fra dette kan man altså se at det er oktober og september der har været de to tørreste måneder

The screenshot shows the OpenRefine interface with a custom text facet applied to the 'months\_no\_water' column. The facet is titled 'months\_no\_water' and shows 11 choices: Oct 74, Sept 70, Nov 51, NULL 45, Aug 33, Dec 11, Jan 2, July 2, Apr 1, June 1, and May 1. The main table displays 131 rows of data. The columns are: no\_group\_count, yes\_group\_count, no\_enough\_water, months\_no\_water, period\_use, exper\_other, other\_meth, res\_change, memb\_assoc, resp\_assoc, and fees\_water. The data shows various counts and categorical values for each row.

| no_group_count | yes_group_count | no_enough_water | months_no_water | period_use | exper_other | other_meth | res_change | memb_assoc | resp_assoc | fees_water |
|----------------|-----------------|-----------------|-----------------|------------|-------------|------------|------------|------------|------------|------------|
| 2              | NULL            | NULL            | NULL            | NULL       | NULL        | NULL       | NULL       | NULL       | NULL       | NULL       |
| NULL           | 3               | yes             | AugSept         | 2          | yes         | no         | NULL       | yes        | no         | no         |
| 1              | NULL            | NULL            | NULL            | NULL       | NULL        | NULL       | NULL       | NULL       | NULL       | NULL       |
| 3              | NULL            | NULL            | NULL            | NULL       | NULL        | NULL       | NULL       | NULL       | NULL       | NULL       |
| 2              | NULL            | NULL            | NULL            | NULL       | NULL        | NULL       | NULL       | NULL       | NULL       | NULL       |
| 1              | NULL            | NULL            | NULL            | NULL       | NULL        | NULL       | NULL       | NULL       | NULL       | NULL       |
| NULL           | 4               | yes             | AugSeptOct      | 10         | yes         | no         | NULL       | no         | NULL       | no         |
| NULL           | 2               | yes             | SeptOct         | 10         | yes         | no         | NULL       | yes        | yes        | no         |
| NULL           | 3               | yes             | OctNov          | 6          | yes         | no         | NULL       | no         | NULL       | no         |

5. Real-Data Challenge: What are the 10 most frequent occupations "erhverv" among unmarried men or women of 20-30 years in [1801 Aarhus](#) census dataset? (hint: first select either men or women to shrink the dataset to a manageable size, then filter by age, and then use merging to cut the erhvervvariation ruthlessly.)

- Vi har valgt at kigge på ugifte kvinder i datasættet og har derfor kun inkluderet de data der er hvor "koen" er markeret som "kvinde" og "civilstand" markeret med "ugift" eller "enke". Da opgaverne er stillet med fokus på alderen 20-30 år, er der altså kun inkluderet data, hvor både "koen" er markeret som "kvinde" og "alder" er markeret med data i intervallet [20;30]. Herefter har vi grupperet dataene, så vi har så få grupperinger, at det er muligt at give et mere overskueligt svar. Grupperingen af dataen er gjort ved

key collision, methaphone 3, og nearest neighbor, levensthein. Her har vi vurderet, hvilken data der var relevant at gruppere. Herefter har vi manuelt grupperet den data, der ikke var mulig at gruppere ved hjælp af Openrefines funktioner, hvilket vi har gjort ved at ændre navnet på den data vi har ville tilføje til en gruppe, nogen af dem havde flere "erhverv" så her valgte vi selektivt hvor de passede bedst ind. Vi har valgt at besvare spørgsmålet med 29 forskellige datasvar, men det ville helt klart være muligt at gruppere dataen yderligere. Dette ville dog indebære en grovere sortering af dataen og dermed også et mere upræcist svar.

- På baggrund af ovenstående kan vi derfor konkludere, at de 10 hyppigste erhverv blandt ugifte kvinder i alderen 20-30 år var:
  - 1) tjenestepige
  - 2) væver
  - 3) ude af stand til at arbejde
  - 4) syerske
  - 5) tjener forældrene
  - 6) nyder almisse
  - 7) spinderske
  - 8) inderste
  - 9) lever af sine midler
  - 10) skræder

Svarene fra Openrefine kan ses i screenshot nedenunder

OpenRefine census 1801 [Permalink](#) Open... Export Help

2256 matching rows (44559 total) Schema Issues Preview Extensions Wikibase

Facet / Filter Undo / Redo 85 / 85 Refresh Reset all Remove all

alder change reset

20 — 30

☒ Numeric ☐ Non-numeric ☒ Blank ☐ Error

erhverv change

value

29 choices Sort by: name count Cluster

tjenestepige 53

væver 33

ude af stand til at arbejde 20

syerske 14

tjener forældrene 14

nyder almisse 12

spinderske 11

inderste 8

lever af sine midler 8

skræder 7

bryggerpige 6

koen change invert reset

2 choices Sort by: name count Cluster

kvinde 2256

mand 2760

|      | ft   | sogn     | amt     | id  | loknr | lokalitet      | bygning | famn | fnavn         | enavn            | koen   | famstand      | alder | civilstand | giftnr | erhverv |
|------|------|----------|---------|-----|-------|----------------|---------|------|---------------|------------------|--------|---------------|-------|------------|--------|---------|
| 399. | 1801 | Borum    | Århus   | 109 | 14    | Borum Bye      |         | 14   | Anne          | Christiansdatter | kvinde | tjenestepige  | 22    | ugift      |        |         |
| 409. | 1801 | Borum    | Århus   | 119 | 15    | Borum Bye      |         | 15   | Anne          | Jensdatter       | kvinde | tjenestepige  | 23    | ugift      |        |         |
| 413. | 1801 | Borum    | Århus   | 123 | 16    | Borum Bye      |         | 16   | Johanne       | Nielsdatter      | kvinde | tjenestepige  | 20    | ugift      |        |         |
| 418. | 1801 | Borum    | Århus   | 128 | 17    | Borum Bye      |         | 17   | Maren         | Larsdatter       | kvinde | tjenestepige  | 21    | ugift      |        |         |
| 425. | 1801 | Borum    | Århus   | 135 | 18    | Borum Bye      |         | 18   | Anne          | Nielsdatter      | kvinde | tjenestepige  | 28    | ugift      |        |         |
| 501. | 1801 | Borum    | Århus   | 211 | 37    | Borum Bye      |         | 37   | Johanne       | Pedersdatter     | kvinde | tjenestepige  | 26    | ugift      |        | væver   |
| 569. | 1801 | Borum    | Århus   | 279 | 55    | Borum Bye      |         | 55   | Ingeborre     | Andersdatter     | kvinde | tjenestepige  | 26    | ugift      |        |         |
| 582. | 1801 | Brabrand | Århus   | 11  | 1     | Brabrand Bye   |         | 1    | Mette Marie   | Andersdatter     | kvinde | tjenestepige  | 20    | ugift      |        |         |
| 601. | 1801 | Brabrand | Århus   | 30  | 4     | Brabrand Bye   |         | 4    | Kirsten       | Nielsdatter      | kvinde | tjenestepige  | 23    | ugift      |        |         |
| 607. | 1801 | Brabrand | Århus   | 36  | 5     | Brabrand Bye   |         | 5    | Lisbeth       | Sørensdatter     | kvinde | tjenestepige  | 21    | ugift      |        |         |
| 610. | 1801 | Brabrand | Århus   | 39  | 6     | Brabrand Bye   |         | 6    | Anne          | Michelsdatter    | kvinde | deres datter  | 24    | ugift      |        |         |
| 728. | 1801 | Brabrand | Århus   | 157 | 33    | Brabrand Bye   |         | 33   | Anne          | Pedersdatter     | kvinde | tjenestepige  | 25    | ugift      |        |         |
| 798. | 1801 | Brabrand | Århus   | 227 | 47    | True Bye       |         | 9    | Mette         | Nielsdatter      | kvinde | konens søster | 20    | ugift      |        |         |
| 807. | 1801 | Brabrand | Århus   | 236 | 49    | True Bye       |         | 11   | Anne          | Jensdatter       | kvinde | hans datter   | 23    | ugift      |        |         |
| 895. | 1801 | Brabrand | Århus   | 324 | 69    | Holmstrupgaard |         | 1    | Anne          | Rasmusdatter     | kvinde | tjenestepige  | 28    | ugift      |        |         |
| 928. | 1801 | Brabrand | Århus   | 357 | 76    | Ydrup Bye      |         | 1    | Anne Kirstine | Sørensdatter     | kvinde | tjenestepige  | 23    | ugift      |        |         |
| 960. | 1801 | Egå      | Randers | 8   | 1     | Eggaee Bye     |         | 1    | Anna          | Madsdatter       | kvinde | tjenestepige  | 28    | ugift      |        |         |
| 974. | 1801 | Egå      | Randers | 22  | 3     | Eggaee Bye     |         | 3    | Mette         | Sørensdatter     | kvinde | tjenestepige  | 21    | ugift      |        |         |
| 978. | 1801 | Egå      | Randers | 26  | 4     | Eggaee Bye     |         | 4    | Karen         | Jørgensdatter    | kvinde | tjenestepige  | 22    | ugift      |        |         |
| 986. | 1801 | Egå      | Randers | 34  | 5     | Eggaee Bye     |         | 5    | Karen         | Olesdatter       | kvinde | tjenestepige  | 29    | ugift      |        |         |
| 994. | 1801 | Egå      | Randers | 42  | 6     | Eggaee Bye     |         | 6    | Anne Kirstine | Pedersdatter     | kvinde | tjenestepige  | 29    | ugift      |        |         |
| 004. | 1801 | Egå      | Randers | 52  | 7     | Eggaee Bye     |         | 7    | Anna Margrete | Emsdatter        | kvinde | tjenestepige  | 21    | ugift      |        |         |
| 017. | 1801 | Egå      | Randers | 65  | 9     | Eggaee Bye     |         | 9    | Karen         | Jensdatter       | kvinde | tjenestepige  | 23    | ugift      |        |         |
| 027. | 1801 | Egå      | Randers | 75  | 11    | Eggaee Bye     |         | 11   | Maren         | Rasmusdatter     | kvinde | tjenestepige  | 20    | ugift      |        |         |

tjenestepige 53

væver 33

ude af stand til at arbejde 20

syerske 14

tjener forældrene 14

nyder almisse 12

spinderske 11

inderste 8

lever af sine midler 8

skræder 7

bryggerpige 6

kokkepige 5

arbejdsløs 2

gaardbeboer 2

har alt frit paa gaarden 2

huusbeboer med jord 2

afdød selveier anders jensens huus med jord 1

amme 1

arbejder med håndarbejde 1

barnepige 1

daglejer 1

gør stoele 1

går med bagerkurven 1

huusmoder 1

i besøg på gården 1

inderste der ellers er udi fast tieneste 1

lever af husnæring 1

mejerske 1

opholdende 1

(blank) 2044