duplicate n-gram detection

Anne Rutten

January 31, 2020

aim:

flag data points that are part of a sequence that is present at least twice in a given data set.

sequence flagging:

- function generates n-grams for specified data, iteratively increasing the sequence length from the specified `min_length` until no more duplicate n-grams are present in the data
- data points that are part of an n-gram that is present in the data more than once are marked with an identifier specific to the sequence

please note:

- the function does not yet check for overlapping sequences within a specific n-gram length, i.e. a sequence “A A B A A B A” will count and mark n-gram “A A B A” as duplicated
- longer sequences will overwrite shorter sequences that they overlap with, i.e., in the above example, 3-gram “A B A” will be overwritten by 4-gram “A A B A”. If “A B A” occurs at a different position as well this may seem an ‘orphan’ sequence (because its brothers got overwritten by the 4-gram). Likewise, shorter n-grams may be only partly overwritten by a longer n-gram.

example usage in Raphaels data:

download from Raphaels github: https://github.com/rroyaute/Royaute-Pruitt-Ecology-2015-Data-and-Code/blob/master/Data%20files/Pardosa_mesocosm_activity_P_R.csv (https://github.com/rroyaute/Royaute-Pruitt-Ecology-2015-Data-and-Code/blob/master/Data%20files/Pardosa_mesocosm_activity_P_R.csv)

1: load data

- enter path to your datafile (csv format)

filename:

/Users/raphael/Dropbox/Research/Royaute&Pruitt_Ecol_Files/Data files/SequenceSniffing/Pardosa_cannibalism_P_R_Ssniff/Pardosa_cannibalism_P_R.csv

get data

raw data:

- view can be expanded

Show 10 **entries** **Search:**

	Replicate_ID	Treatment	Phase	Pardosa	duplicatedRow
1	1	Inactive	Initial	8	false
2	2	Inactive	Initial	8	false
3	3	Inactive	Initial	8	false
4	4	Inactive	Initial	8	false
5	5	Inactive	Initial	8	false
6	6	Inactive	Initial	8	false
7	7	Inactive	Initial	8	false
8	8	Inactive	Initial	8	false
9	9	Inactive	Initial	8	false
10	10	Inactive	Initial	8	false

Showing 1 to 10 of 90 entries Previous 1 2 3 4 5 … 9 Next

2: controls:

- `focal column range start` : start of column range (including)
- `focal column range end` : end of column range (including)
- `min sequence length` : detect recurring sequences of this length or greater
- `ignore repeats of same value` : in some cases (default values, censored data) many repeat sequences are expected. Check this box to ignore sequences consisting of the same value, repeated.

focal column range start:

4

focal column range end:

4

min sequence length

4

□ ignore repeats of same value

detect duplicates

data summary:

- `key` : focal column name
- `cardinality` : number of distinct values in the data
- `max_rle` : longest sequence of repeats of the same number (if high, you may want to select `ignore repeats of same value` above)
- `max_rle_at_level` : the value(s) that is/are repeated `max_rle` times
- `n_duplicates` : number of datapoints that are part of any non-unique sequence of length > `min sequence length`
- `n_total` : number of rows
- `fraction` : `n_duplicates` / `n_total`

Show 10 **entries** **Search:**

	key	cardinality	max_rle	max_rle_at_level	n_duplicates	n_total	fraction
1	Pardosa	5	45	8	65	90	0.72

Showing 1 to 1 of 1 entries Previous 1 Next

detected sequences:

- view can be expanded
- colours: sequence in column is not unique
- grey: row is not unique
- `value_*` columns: the original data for the focal columna
- `ngramID_*` columns: ngramID of the sequence in the corresponding `value_*` column

`ngramID` is generated as follows: n+ `sequence length` + `Sequence ID`. `Sequence ID` does not carry more information.

Download csv (session/01d25717a968839ee1c5d5ab38c7dc85/download/downloadData?w=)

Show 100 **entries** **Search:**

	Replicate_ID	Treatment	Phase	duplicatedRow	originalRowID	value_Pardosa	ngramID_Pardosa
1	1	Inactive	Initial	false	1	8	n44Seq1
2	2	Inactive	Initial	false	2	8	n44Seq1
3	3	Inactive	Initial	false	3	8	n44Seq1
4	4	Inactive	Initial	false	4	8	n44Seq1
5	5	Inactive	Initial	false	5	8	n44Seq1
6	6	Inactive	Initial	false	6	8	n44Seq1
7	7	Inactive	Initial	false	7	8	n44Seq1
8	8	Inactive	Initial	false	8	8	n44Seq1

	Replicate_ID	Treatment	Phase	duplicatedRow	originalRowID	value_Pardosa	ngramID_Pardosa
9	9	Inactive	Initial	false	9	8	n44Seq1
10	10	Inactive	Initial	false	10	8	n44Seq1
11	11	Inactive	Initial	false	11	8	n44Seq1
12	12	Inactive	Initial	false	12	8	n44Seq1
13	13	Inactive	Initial	false	13	8	n44Seq1
14	14	Inactive	Initial	false	14	8	n44Seq1
15	15	Inactive	Initial	false	15	8	n44Seq1
16	16	Active	Initial	false	16	8	n44Seq1
17	17	Active	Initial	false	17	8	n44Seq1
18	18	Active	Initial	false	18	8	n44Seq1
19	19	Active	Initial	false	19	8	n44Seq1
20	20	Active	Initial	false	20	8	n44Seq1
21	21	Active	Initial	false	21	8	n44Seq1
22	22	Active	Initial	false	22	8	n44Seq1
23	23	Active	Initial	false	23	8	n44Seq1
24	24	Active	Initial	false	24	8	n44Seq1
25	25	Active	Initial	false	25	8	n44Seq1
26	26	Active	Initial	false	26	8	n44Seq1
27	27	Active	Initial	false	27	8	n44Seq1
28	28	Active	Initial	false	28	8	n44Seq1
29	29	Active	Initial	false	29	8	n44Seq1
30	30	Active	Initial	false	30	8	n44Seq1
31	31	Mixed	Initial	false	31	8	n44Seq1
32	32	Mixed	Initial	false	32	8	n44Seq1
33	33	Mixed	Initial	false	33	8	n44Seq1
34	34	Mixed	Initial	false	34	8	n44Seq1
35	35	Mixed	Initial	false	35	8	n44Seq1
36	36	Mixed	Initial	false	36	8	n44Seq1
37	37	Mixed	Initial	false	37	8	n44Seq1
38	38	Mixed	Initial	false	38	8	n44Seq1
39	39	Mixed	Initial	false	39	8	n44Seq1
40	40	Mixed	Initial	false	40	8	n44Seq1
41	41	Mixed	Initial	false	41	8	n44Seq1
42	42	Mixed	Initial	false	42	8	n44Seq1

	Replicate_ID	Treatment	Phase	duplicatedRow	originalRowID	value_Pardosa	ngramID_Pardosa
43	43	Mixed	Initial	false	43	8	n44Seq1
44	44	Mixed	Initial	false	44	8	n44Seq1
45	45	Mixed	Initial	false	45	8	n44Seq1
46	1	Inactive	Post-Treatment	false	46	5	n5Seq2
47	2	Inactive	Post-Treatment	false	47	6	n5Seq2
48	3	Inactive	Post-Treatment	false	48	5	n5Seq2
49	4	Inactive	Post-Treatment	false	49	7	n5Seq2
50	5	Inactive	Post-Treatment	false	50	7	n5Seq2
51	6	Inactive	Post-Treatment	false	51	6	n4Seq3
52	7	Inactive	Post-Treatment	false	52	7	n4Seq3
53	8	Inactive	Post-Treatment	false	53	5	
54	9	Inactive	Post-Treatment	false	54	7	
55	10	Inactive	Post-Treatment	false	55	8	
56	11	Inactive	Post-Treatment	false	56	6	
57	12	Inactive	Post-Treatment	false	57	5	
58	13	Inactive	Post-Treatment	false	58	6	
59	14	Inactive	Post-Treatment	false	59	6	
60	15	Inactive	Post-Treatment	false	60	6	
61	16	Active	Post-Treatment	false	61	5	
62	17	Active	Post-Treatment	false	62	6	
63	18	Active	Post-Treatment	false	63	5	
64	19	Active	Post-Treatment	false	64	8	
65	20	Active	Post-Treatment	false	65	4	

	Replicate_ID	Treatment	Phase	duplicatedRow	originalRowID	value_Pardosa	ngramID_Pardosa
66	21	Active	Post-Treatment	false	66	5	n4Seq2
67	22	Active	Post-Treatment	false	67	6	n4Seq2
68	23	Active	Post-Treatment	false	68	5	n4Seq2
69	24	Active	Post-Treatment	false	69	7	n4Seq2
70	25	Active	Post-Treatment	false	70	5	n5Seq2
71	26	Active	Post-Treatment	false	71	6	n5Seq2
72	27	Active	Post-Treatment	false	72	5	n5Seq2
73	28	Active	Post-Treatment	false	73	7	n5Seq2
74	29	Active	Post-Treatment	false	74	7	n5Seq2
75	30	Active	Post-Treatment	false	75	5	
76	31	Mixed	Post-Treatment	false	76	8	
77	32	Mixed	Post-Treatment	false	77	6	
78	33	Mixed	Post-Treatment	false	78	7	
79	34	Mixed	Post-Treatment	false	79	6	
80	35	Mixed	Post-Treatment	false	80	6	
81	36	Mixed	Post-Treatment	false	81	5	
82	37	Mixed	Post-Treatment	false	82	5	
83	38	Mixed	Post-Treatment	false	83	6	
84	39	Mixed	Post-Treatment	false	84	7	n4Seq3
85	40	Mixed	Post-Treatment	false	85	7	n4Seq3
86	41	Mixed	Post-Treatment	false	86	6	n4Seq3
87	42	Mixed	Post-Treatment	false	87	7	n4Seq3

	Replicate_ID	Treatment	Phase	duplicatedRow	originalRowID	value_Pardosa	ngramID_Pardosa
88	43	Mixed	Post-Treatment	false	88	7	
89	44	Mixed	Post-Treatment	false	89	7	
90	45	Mixed	Post-Treatment	false	90	7	

Showing 1 to 90 of 90 entries

Previous 1 Next

3: randomisation controls:

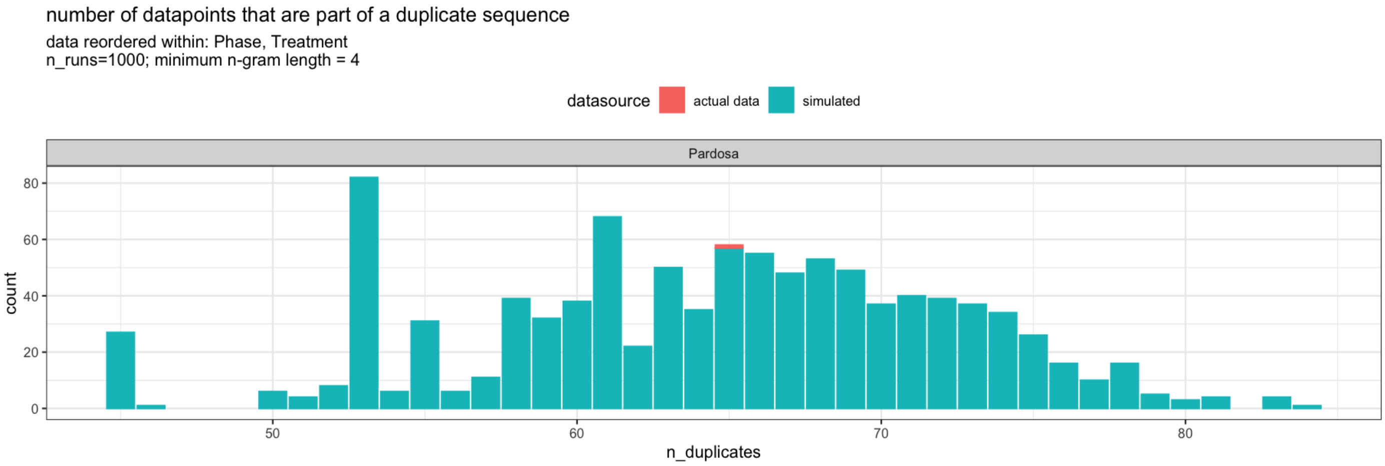
- data will be reordered per column within the grouping levels specified
- **note:** *the order of the selected fields should reflect the data structure.*

reorder within:

Phase Treatment

run random reordering

randomisation result:


number of datapoints that are part of a duplicate sequence
data reordered within: Phase, Treatment
n_runs=1000; minimum n-gram length = 4
datasource
actual data
simulated
Pardosa
count
n_duplicates