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:

files/SequenceSniffing/Pardosa_predation_single_Ssniff/Pardosa_predation_single_v2.csv

get data

raw data:

- view can be expanded

Show 100 ⬍ **entries** **Search:**

	Prey	Activity	Capture	duplicatedRow
1	Blister_beetle	7	1	false
2	Blister_beetle	16	0	true
3	Blister_beetle	3	1	false
4	Blister_beetle	13	1	false
5	Blister_beetle	22	0	false
6	Blister_beetle	5	0	true
7	Blister_beetle	18	1	true
8	Blister_beetle	20	1	false
9	Blister_beetle	16	1	true
10	Blister_beetle	12	0	false
11	Blister_beetle	18	1	true
12	Blister_beetle	5	1	true
13	Leaf_hopper	8	1	true
14	Leaf_hopper	17	1	true

	Prey	Activity	Capture	duplicatedRow
15	Leaf_hopper	25	1	false
16	Leaf_hopper	8	1	true
17	Leaf_hopper	17	1	true
18	Leaf_hopper	22	1	false
19	Leaf_hopper	19	1	false
20	Leaf_hopper	8	1	true
21	Leaf_hopper	21	0	false
22	Leaf_hopper	8	1	true
23	Leaf_hopper	9	1	false
24	Leaf_hopper	8	1	true
25	Beet_armyworm	20	0	false
26	Beet_armyworm	6	0	false
27	Beet_armyworm	23	0	true
28	Beet_armyworm	17	1	false
29	Beet_armyworm	19	0	true
30	Beet_armyworm	13	0	false
31	Beet_armyworm	23	0	true
32	Beet_armyworm	19	0	true
33	Beet_armyworm	21	1	true
34	Beet_armyworm	21	1	true
35	Beet_armyworm	9	0	false
36	Beet_armyworm	18	0	false
37	Pea_aphid	17	0	true
38	Pea_aphid	24	0	true
39	Pea_aphid	15	0	true
40	Pea_aphid	6	0	false
41	Pea_aphid	5	0	false
42	Pea_aphid	15	0	true
43	Pea_aphid	3	0	false
44	Pea_aphid	19	0	false
45	Pea_aphid	8	0	false
46	Pea_aphid	17	1	true
47	Pea_aphid	18	0	true
48	Pea_aphid	18	0	true
49	Pea_aphid	11	1	false
50	Pea_aphid	24	0	true
51	Sharpshooter	24	1	false
52	Sharpshooter	28	1	false

	Prey	Activity	Capture	duplicatedRow
53	Sharpshooter	18	1	true
54	Sharpshooter	1	1	false
55	Sharpshooter	19	1	false
56	Sharpshooter	18	1	true
57	Sharpshooter	17	1	true
58	Sharpshooter	13	0	false
59	Sharpshooter	8	1	true
60	Sharpshooter	17	0	true
61	Sharpshooter	8	1	true
62	Sharpshooter	11	1	true
63	Sharpshooter	15	1	false
64	Sharpshooter	11	1	true
65	Alfalfa_weevil	18	0	false
66	Alfalfa_weevil	11	0	true
67	Alfalfa_weevil	11	0	true
68	Alfalfa_weevil	5	1	false
69	Alfalfa_weevil	17	0	true
70	Alfalfa_weevil	20	0	false
71	Alfalfa_weevil	21	0	true
72	Alfalfa_weevil	17	0	true
73	Alfalfa_weevil	19	0	false
74	Alfalfa_weevil	21	0	true
75	Alfalfa_weevil	8	1	false
76	Alfalfa_weevil	11	0	true
77	Alfalfa_weevil	15	1	false
78	Alfalfa_weevil	11	1	true

Showing 1 to 78 of 78 entries

Previous 1 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: 2

focal column range end: 3

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	Activity	22	2	21,18,24,11	8	78	0.1
2	Capture	2	11	0	78	78	1

Showing 1 to 2 of 2 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:**

	Prey	duplicatedRow	originalRowID	value_Activity	value_Capture	ngramID_Activity	ngramID_Captu
1	Blister_beetle	false	1	7	1		n4Seq3
2	Blister_beetle	true	2	16	0		n5Seq14
3	Blister_beetle	false	3	3	1		n5Seq14
4	Blister_beetle	false	4	13	1		n5Seq14
5	Blister_beetle	false	5	22	0		n5Seq14
6	Blister_beetle	true	6	5	0		n5Seq17
7	Blister_beetle	true	7	18	1		n7Seq6
8	Blister_beetle	false	8	20	1		n7Seq6
9	Blister_beetle	true	9	16	1		n9Seq6
10	Blister_beetle	false	10	12	0		n9Seq6
11	Blister_beetle	true	11	18	1		n9Seq7
12	Blister_beetle	true	12	5	1		n9Seq7
13	Leaf_hopper	true	13	8	1		n9Seq7
14	Leaf_hopper	true	14	17	1		n9Seq7
15	Leaf_hopper	false	15	25	1		n9Seq7
16	Leaf_hopper	true	16	8	1		n9Seq7
17	Leaf_hopper	true	17	17	1		n9Seq7
18	Leaf_hopper	false	18	22	1		n9Seq7
19	Leaf_hopper	false	19	19	1		n9Seq7
20	Leaf_hopper	true	20	8	1		n9Seq7
21	Leaf_hopper	false	21	21	0		n9Seq5
22	Leaf_hopper	true	22	8	1		n11Seq1
23	Leaf_hopper	false	23	9	1		n11Seq1

	Prey	duplicatedRow	originalRowID	value_Activity	value_Capture	ngramID_Activity	ngramID_Captu
24	Leaf_hopper	true	24	8	1		n11Seq1
25	Beet_armyworm	false	25	20	0		n11Seq1
26	Beet_armyworm	false	26	6	0		n11Seq1
27	Beet_armyworm	true	27	23	0		n11Seq1
28	Beet_armyworm	false	28	17	1		n11Seq1
29	Beet_armyworm	true	29	19	0		n11Seq1
30	Beet_armyworm	false	30	13	0		n11Seq1
31	Beet_armyworm	true	31	23	0		n11Seq1
32	Beet_armyworm	true	32	19	0		n11Seq1
33	Beet_armyworm	true	33	21	1		n5Seq14
34	Beet_armyworm	true	34	21	1		n7Seq8
35	Beet_armyworm	false	35	9	0		n10Seq2
36	Beet_armyworm	false	36	18	0		n10Seq2
37	Pea_aphid	true	37	17	0		n10Seq2
38	Pea_aphid	true	38	24	0		n10Seq2
39	Pea_aphid	true	39	15	0		n10Seq2
40	Pea_aphid	false	40	6	0		n10Seq2
41	Pea_aphid	false	41	5	0		n10Seq2
42	Pea_aphid	true	42	15	0		n10Seq2
43	Pea_aphid	false	43	3	0		n10Seq2
44	Pea_aphid	false	44	19	0		n10Seq2
45	Pea_aphid	false	45	8	0		n10Seq2
46	Pea_aphid	true	46	17	1		n8Seq9
47	Pea_aphid	true	47	18	0		n8Seq9
48	Pea_aphid	true	48	18	0		n7Seq12
49	Pea_aphid	false	49	11	1		n9Seq6
50	Pea_aphid	true	50	24	0		n9Seq6
51	Sharpshooter	false	51	24	1		n9Seq6
52	Sharpshooter	false	52	28	1		n9Seq6
53	Sharpshooter	true	53	18	1		n9Seq6
54	Sharpshooter	false	54	1	1		n9Seq6
55	Sharpshooter	false	55	19	1		n9Seq6
56	Sharpshooter	true	56	18	1		n9Seq6
57	Sharpshooter	true	57	17	1		n9Seq6
58	Sharpshooter	false	58	13	0		n9Seq5
59	Sharpshooter	true	59	8	1		n9Seq5
60	Sharpshooter	true	60	17	0		n7Seq12
61	Sharpshooter	true	61	8	1	n4Seq15	n7Seq12

	Prey	duplicatedRow	originalRowID	value_Activity	value_Capture	ngramID_Activity	ngramID_Captu
62	Sharpshooter	true	62	11	1	n4Seq15	n11Seq1
63	Sharpshooter	false	63	15	1	n4Seq15	n11Seq1
64	Sharpshooter	true	64	11	1	n4Seq15	n11Seq1
65	Alfalfa_weevil	false	65	18	0		n11Seq1
66	Alfalfa_weevil	true	66	11	0		n11Seq1
67	Alfalfa_weevil	true	67	11	0		n11Seq1
68	Alfalfa_weevil	false	68	5	1		n11Seq1
69	Alfalfa_weevil	true	69	17	0		n11Seq1
70	Alfalfa_weevil	false	70	20	0		n11Seq1
71	Alfalfa_weevil	true	71	21	0		n11Seq1
72	Alfalfa_weevil	true	72	17	0		n11Seq1
73	Alfalfa_weevil	false	73	19	0		n8Seq9
74	Alfalfa_weevil	true	74	21	0		n8Seq9
75	Alfalfa_weevil	false	75	8	1	n4Seq15	n8Seq9
76	Alfalfa_weevil	true	76	11	0	n4Seq15	n8Seq9
77	Alfalfa_weevil	false	77	15	1	n4Seq15	n6Seq16
78	Alfalfa_weevil	true	78	11	1	n4Seq15	n6Seq16

Showing 1 to 78 of 78 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:

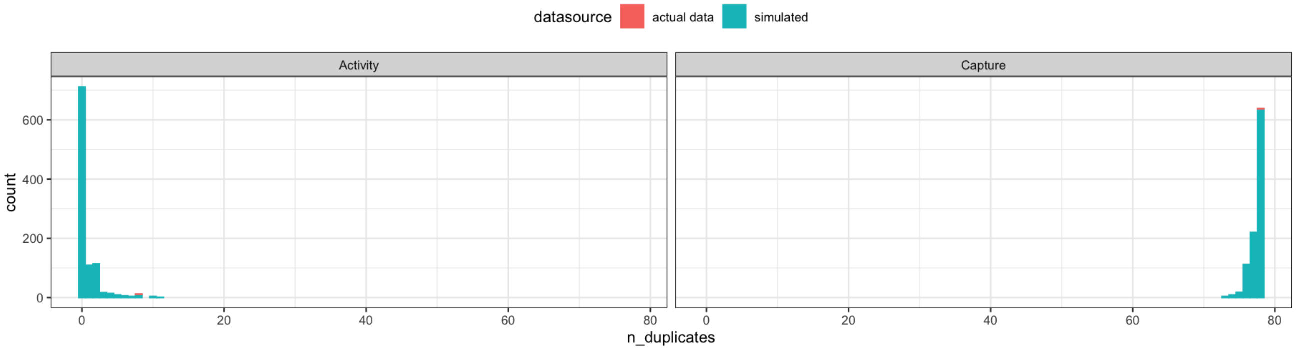
Prey

run random reordering

randomisation result:

number of datapoints that are part of a duplicate sequence

data reordered within: Prey
n_runs=1000; minimum n-gram length = 4


datasource
🟥 actual data
🟩 simulated
Activity
Capture
count
n_duplicates