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_mesocosm_activity_P_R_Ssniff/Pardosa_mesocosm_activity_P_R.csv

get data

raw data:

- view can be expanded

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

	Replicate_ID	Treatment	Pardosa_1	Pardosa_2	Pardosa_3	Pardosa_4	Pardosa_5	Pardosa_6	Pardosa_7
1	1	Inactive	7	7	8	8	4	2	8
2	2	Inactive	8	6	11	6	9	8	7
3	3	Inactive	4	8	13	8	8	3	8
4	4	Inactive	7	11	12	6	11	7	6
5	5	Inactive	8	5	13	3	13	4	8
6	6	Inactive	12	7	7	8	7	8	7
7	7	Inactive	7	4	2	6	11	5	12
8	8	Inactive	5	9	8	7	8	7	11
9	9	Inactive	8	8	3	5	13	5	8
10	10	Inactive	6	11	7	8	13	3	7
11	11	Inactive	7	13	4	12	7	8	11
12	12	Inactive	8	7	8	9	2	6	12
13	13	Inactive	9	6	5	7	8	7	8
14	14	Inactive	7	8	7	8	3	5	9
15	15	Inactive	11	11	5	7	9	11	7
16	16	Active	13	21	14	23	14	16	17
17	17	Active	14	32	21	18	21	19	21

	Replicate_ID	Treatment	Pardosa_1	Pardosa_2	Pardosa_3	Pardosa_4	Pardosa_5	Pardosa_6	Pardosa_7
18	18	Active	18	25	15	26	15	18	23
19	19	Active	16	21	17	18	17	16	28
20	20	Active	21	24	21	16	21	21	31
21	21	Active	23	18	23	19	23	20	17
22	22	Active	22	16	28	18	28	27	18
23	23	Active	28	19	31	16	13	15	16
24	24	Active	16	18	17	21	15	16	17
25	25	Active	17	16	18	20	17	17	15
26	26	Active	18	21	16	27	18	21	22
27	27	Active	15	20	17	28	21	22	21
28	28	Active	19	27	15	23	29	18	22
29	29	Active	20	28	21	19	17	21	23
30	30	Active	21	23	18	17	23	28	26
31	31	Mixed	8	8	7	8	19	14	28
32	32	Mixed	3	7	5	11	18	21	16
33	33	Mixed	7	11	13	3	16	15	17
34	34	Mixed	4	13	6	8	21	17	18
35	35	Mixed	8	9	7	11	20	21	15
36	36	Mixed	5	7	8	7	27	23	19
37	37	Mixed	7	8	13	5	28	28	20
38	38	Mixed	5	12	7	8	23	31	23
39	39	Mixed	3	8	11	12	19	17	25
40	40	Mixed	8	13	12	9	17	18	27
41	41	Mixed	6	9	10	7	13	18	21
42	42	Mixed	7	7	9	8	15	16	17
43	43	Mixed	5	13	7	7	17	21	19
44	44	Mixed	11	9	8	12	18	20	23
45	45	Mixed	13	11	11	13	21	17	23

Showing 1 to 45 of 45 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: 3

focal column range end: 10

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_1	21	1		22	45	0.49
2	Pardosa_2	20	1		32	45	0.71
3	Pardosa_3	21	1		24	45	0.53
4	Pardosa_4	19	1		31	45	0.69
5	Pardosa_5	20	2	13	38	45	0.84
6	Pardosa_6	21	2	18	36	45	0.8
7	Pardosa_7	20	2	23	17	45	0.38
8	Pardosa_8	19	1		15	45	0.33

Showing 1 to 8 of 8 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	mean_activity	duplicatedRow	originalRowID	value_Pardosa_1	value_Pardosa_2	valu
1	1	Inactive	6.5	false	1	7	7	
2	2	Inactive	7.75	false	2	8	6	
3	3	Inactive	7.5	false	3	4	8	
4	4	Inactive	8.375	false	4	7	11	
5	5	Inactive	8.125	false	5	8	5	
6	6	Inactive	8.625	false	6	12	7	
7	7	Inactive	7	false	7	7	4	
8	8	Inactive	7.75	false	8	5	9	
9	9	Inactive	7.25	false	9	8	8	
10	10	Inactive	8.375	false	10	6	11	
11	11	Inactive	8.75	false	11	7	13	
12	12	Inactive	7.25	false	12	8	7	
13	13	Inactive	7.375	false	13	9	6	
14	14	Inactive	6.75	false	14	7	8	
15	15	Inactive	8.625	false	15	11	11	
16	16	Active	17	false	16	13	21	
17	17	Active	20.375	false	17	14	32	
18	18	Active	20.125	false	18	18	25	
19	19	Active	19.5	false	19	16	21	
20	20	Active	22.875	false	20	21	24	

	Replicate_ID	Treatment	mean_activity	duplicatedRow	originalRowID	value_Pardosa_1	value_Pardosa_2	valu
21	21	Active	21.75	false	21	23	18	
22	22	Active	21.75	false	22	22	16	
23	23	Active	20	false	23	28	19	
24	24	Active	17	false	24	16	18	
25	25	Active	17.875	false	25	17	16	
26	26	Active	20.125	false	26	18	21	
27	27	Active	20.625	false	27	15	20	
28	28	Active	22.125	false	28	19	27	
29	29	Active	21.375	false	29	20	28	
30	30	Active	22	false	30	21	23	
31	31	Mixed	14.375	false	31	8	8	
32	32	Mixed	12.75	false	32	3	7	
33	33	Mixed	13.625	false	33	7	11	
34	34	Mixed	13.625	false	34	4	13	
35	35	Mixed	13.5	false	35	8	9	
36	36	Mixed	14.25	false	36	5	7	
37	37	Mixed	15.625	false	37	7	8	
38	38	Mixed	16.25	false	38	5	12	
39	39	Mixed	15.125	false	39	3	8	
40	40	Mixed	15.625	false	40	8	13	
41	41	Mixed	13	false	41	6	9	
42	42	Mixed	11.875	false	42	7	7	
43	43	Mixed	13.25	false	43	5	13	
44	44	Mixed	14.25	false	44	11	9	
45	45	Mixed	15.875	false	45	13	11	

Showing 1 to 45 of 45 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:

Treatment

run random reordering

randomisation result:

number of datapoints that are part of a duplicate sequence

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