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/Mesocosms_data_Ssniff/Mesocosms_data.csv

get data

raw data:

- view can be expanded

Show 10 entries Search:

	Replicate.ID	Replicate_ID2	Treatment	Phase	Blister_beetle	Leaf_hoppers	Beet_armyworm	Pea_aphid
1	1	B1	Inactive	Post-Treatment	12	8	10	55
2	2	B2	Inactive	Post-Treatment	13	9	9	27
3	3	B3	Inactive	Post-Treatment	12	7	6	42
4	4	B4	Inactive	Post-Treatment	11	9	8	33
5	5	B5	Inactive	Post-Treatment	10	8	7	29
6	6	B6	Inactive	Post-Treatment	7	9	10	56
7	7	B7	Inactive	Post-Treatment	8	11	10	49
8	8	B8	Inactive	Post-Treatment	9	12	7	59
9	9	B9	Inactive	Post-Treatment	7	13	8	55
10	10	B10	Inactive	Post-Treatment	10	8	6	67

Showing 1 to 10 of 55 entries Previous 1 2 3 4 5 6 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: 5

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	Alfalfa_weevil	9	2	6,4,2,5	20	55	0.36
2	Beet_armyworm	7	2	10,7	30	55	0.55
3	Blister_beetle	13	2	2	4	55	0.07
4	Leaf_hoppers	16	2	2	17	55	0.31
5	Pea_aphid	37	1		0	55	0
6	Sharpshooter	7	2	2	47	55	0.85

Showing 1 to 6 of 6 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	Replicate_ID2	Treatment	Phase	duplicatedRow	originalRowID	value_Blister_beetle	value_Leaf
1	1	B1	Inactive	Post-Treatment	false	1	12	
2	2	B2	Inactive	Post-Treatment	false	2	13	
3	3	B3	Inactive	Post-Treatment	false	3	12	
4	4	B4	Inactive	Post-Treatment	false	4	11	
5	5	B5	Inactive	Post-Treatment	false	5	10	
6	6	B6	Inactive	Post-Treatment	false	6	7	
7	7	B7	Inactive	Post-Treatment	false	7	8	

	Replicate.ID	Replicate_ID2	Treatment	Phase	duplicatedRow	originalRowID	value_Blister_beetle	value_Leaf
8	8	B8	Inactive	Post-Treatment	false	8	9	
9	9	B9	Inactive	Post-Treatment	false	9	7	
10	10	B10	Inactive	Post-Treatment	false	10	10	
11	11	B11	Inactive	Post-Treatment	false	11	11	
12	12	B12	Inactive	Post-Treatment	false	12	12	
13	13	B13	Inactive	Post-Treatment	false	13	9	
14	14	B14	Inactive	Post-Treatment	false	14	8	
15	15	B15	Inactive	Post-Treatment	false	15	9	
16	1	C1	Active	Post-Treatment	false	16	13	
17	2	C2	Active	Post-Treatment	false	17	14	
18	3	C3	Active	Post-Treatment	false	18	12	
19	4	C4	Active	Post-Treatment	false	19	13	
20	5	C5	Active	Post-Treatment	false	20	14	
21	6	C6	Active	Post-Treatment	false	21	10	
22	7	C7	Active	Post-Treatment	false	22	6	
23	8	C8	Active	Post-Treatment	false	23	10	
24	9	C9	Active	Post-Treatment	false	24	8	
25	10	C10	Active	Post-Treatment	false	25	9	
26	11	C11	Active	Post-Treatment	false	26	12	
27	12	C12	Active	Post-Treatment	false	27	9	
28	13	C13	Active	Post-Treatment	false	28	5	
29	14	C14	Active	Post-Treatment	false	29	11	
30	15	C15	Active	Post-Treatment	false	30	6	
31	1	D1	Mixed	Post-Treatment	false	31	11	
32	2	D2	Mixed	Post-Treatment	false	32	7	
33	3	D3	Mixed	Post-Treatment	false	33	5	

	Replicate.ID	Replicate_ID2	Treatment	Phase	duplicatedRow	originalRowID	value_Blister_beetle	value_Leaf
34	4	D4	Mixed	Post-Treatment	false	34	8	
35	5	D5	Mixed	Post-Treatment	false	35	6	
36	6	D6	Mixed	Post-Treatment	false	36	2	
37	7	D7	Mixed	Post-Treatment	false	37	9	
38	8	D8	Mixed	Post-Treatment	false	38	11	
39	9	D9	Mixed	Post-Treatment	false	39	2	
40	10	D10	Mixed	Post-Treatment	false	40	10	
41	11	D11	Mixed	Post-Treatment	false	41	4	
42	12	D12	Mixed	Post-Treatment	false	42	7	
43	13	D13	Mixed	Post-Treatment	false	43	6	
44	14	D14	Mixed	Post-Treatment	false	44	2	
45	15	D15	Mixed	Post-Treatment	false	45	2	
46	1	A1	Control	Post-Treatment	false	46	12	
47	2	A2	Control	Post-Treatment	false	47	13	
48	3	A3	Control	Post-Treatment	false	48	12	
49	4	A4	Control	Post-Treatment	false	49	13	
50	5	A5	Control	Post-Treatment	false	50	16	
51	6	A6	Control	Post-Treatment	false	51	10	
52	7	A7	Control	Post-Treatment	false	52	13	
53	8	A8	Control	Post-Treatment	false	53	14	
54	9	A9	Control	Post-Treatment	false	54	13	
55	10	A10	Control	Post-Treatment	false	55	11	

Showing 1 to 55 of 55 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:

Replicate.ID

run random reordering

randomisation result: