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_mesocosm_activity_P_R_Ssniff/Pardosa_mesocosm_activity_P_R_wide.csv

get data

raw data:

- view can be expanded

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

	Spider_ID	R_1	R_2	R_3	R_4	R_5	R_6	R_7	R_8	R_9	R_10	R_11	R_12	R_13	R_
1	Pardosa_1	7	8	4	7	8	12	7	5	8	6	7	8	9	
2	Pardosa_2	7	6	8	11	5	7	4	9	8	11	13	7	6	
3	Pardosa_3	8	11	13	12	13	7	2	8	3	7	4	8	5	
4	Pardosa_4	8	6	8	6	3	8	6	7	5	8	12	9	7	
5	Pardosa_5	4	9	8	11	13	7	11	8	13	13	7	2	8	
6	Pardosa_6	2	8	3	7	4	8	5	7	5	3	8	6	7	
7	Pardosa_7	8	7	8	6	8	7	12	11	8	7	11	12	8	
8	Pardosa_8	8	7	8	7	11	13	9	7	8	12	8	6	9	
9	Pardosa_1	13	14	18	16	21	23	22	28	16	17	18	15	19	
10	Pardosa_2	21	32	25	21	24	18	16	19	18	16	21	20	27	
11	Pardosa_3	14	21	15	17	21	23	28	31	17	18	16	17	15	
12	Pardosa_4	23	18	26	18	16	19	18	16	21	20	27	28	23	
13	Pardosa_5	14	21	15	17	21	23	28	13	15	17	18	21	29	
14	Pardosa_6	16	19	18	16	21	20	27	15	16	17	21	22	18	
15	Pardosa_7	17	21	23	28	31	17	18	16	17	15	22	21	22	
16	Pardosa_8	18	17	21	23	28	31	17	22	16	23	18	21	24	
17	Pardosa_1	8	3	7	4	8	5	7	5	3	8	6	7	5	

	Spider_ID	R_1	R_2	R_3	R_4	R_5	R_6	R_7	R_8	R_9	R_10	R_11	R_12	R_13	R_
18	Pardosa_2	8	7	11	13	9	7	8	12	8	13	9	7	13	
19	Pardosa_3	7	5	13	6	7	8	13	7	11	12	10	9	7	
20	Pardosa_4	8	11	3	8	11	7	5	8	12	9	7	8	7	
21	Pardosa_5	19	18	16	21	20	27	28	23	19	17	13	15	17	
22	Pardosa_6	14	21	15	17	21	23	28	31	17	18	18	16	21	
23	Pardosa_7	28	16	17	18	15	19	20	23	25	27	21	17	19	
24	Pardosa_8	23	21	27	22	17	18	16	21	26	21	20	16	17	

Showing 1 to 24 of 24 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: 16

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	R_1	13	2	7,8	0	24	0
2	R_10	16	2	17	0	24	0
3	R_11	15	1		0	24	0
4	R_12	13	2	21,7	8	24	0.33
5	R_13	16	2	7	6	24	0.25
6	R_14	16	1		4	24	0.17
7	R_15	13	2	11	0	24	0
8	R_2	14	2	7	0	24	0
9	R_3	15	2	8	0	24	0
10	R_4	14	1		0	24	0

Showing 1 to 10 of 15 entries Previous 1 2 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:

	Spider_ID	Treatment	duplicatedRow	originalRowID	value_R_1	value_R_2	value_R_3	value_R_4	value_R
1	Pardosa_1	Inactive	false	1	7	8	4	7	
2	Pardosa_2	Inactive	false	2	7	6	8	11	
3	Pardosa_3	Inactive	false	3	8	11	13	12	
4	Pardosa_4	Inactive	false	4	8	6	8	6	
5	Pardosa_5	Inactive	false	5	4	9	8	11	
6	Pardosa_6	Inactive	false	6	2	8	3	7	
7	Pardosa_7	Inactive	false	7	8	7	8	6	
8	Pardosa_8	Inactive	false	8	8	7	8	7	
9	Pardosa_1	Active	false	9	13	14	18	16	
10	Pardosa_2	Active	false	10	21	32	25	21	
11	Pardosa_3	Active	true	11	14	21	15	17	
12	Pardosa_4	Active	false	12	23	18	26	18	
13	Pardosa_5	Active	false	13	14	21	15	17	
14	Pardosa_6	Active	false	14	16	19	18	16	
15	Pardosa_7	Active	false	15	17	21	23	28	
16	Pardosa_8	Active	false	16	18	17	21	23	
17	Pardosa_1	Mixed	false	17	8	3	7	4	
18	Pardosa_2	Mixed	false	18	8	7	11	13	
19	Pardosa_3	Mixed	false	19	7	5	13	6	
20	Pardosa_4	Mixed	false	20	8	11	3	8	
21	Pardosa_5	Mixed	false	21	19	18	16	21	
22	Pardosa_6	Mixed	true	22	14	21	15	17	
23	Pardosa_7	Mixed	false	23	28	16	17	18	
24	Pardosa_8	Mixed	false	24	23	21	27	22	

Showing 1 to 24 of 24 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
datasource
actual data
simulated
R_1
R_10
R_11
R_12
R_13
R_14
R_15
R_2
R_3
R_4
R_5
R_6
R_7
R_8
R_9
count
n_duplicates