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/Repeatability_P_R_Ssniff/Repeatability_wide_P_R.csv

get data

raw data:

- view can be expanded

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

	Pardosa_ID	R_1	R_2	R_3	R_4	R_5	R_6	R_7	R_8	Size	duplicatedRow
1	1	14	20	21	16	20	23	17	18	1.4	false
2	2	21	16	17	24	17	21	28	21	1.2	false
3	3	15	11	15	16	12	19	18	11	1.5	false
4	4	17	4	9	14	5	23	24	17	1.3	false
5	5	21	7	20	21	25	21	18	17	1.3	false
6	6	23	14	15	16	17	18	19	21	1.7	false
7	7	8	21	24	25	13	9	15	8	1.6	false
8	8	6	2	7	8	16	21	6	8	1.2	false
9	9	12	14	6	7	22	8	3	11	1.4	false

	Pardosa_ID	R_1	R_2	R_3	R_4	R_5	R_6	R_7	R_8	Size	duplicatedRow
10	10	13	16	17	7	5	20	19	15	1.3	false

Showing 1 to 10 of 19 entries Previous 1 2 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: 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	R_1	11	1		0	19	0
2	R_2	14	1		0	19	0
3	R_3	13	2	17	0	19	0
4	R_4	12	2	7	0	19	0
5	R_5	13	1		0	19	0
6	R_6	11	1		0	19	0
7	R_7	12	1		0	19	0
8	R_8	11	2	17,8,11	0	19	0
9	Size	9	2	1.3,1.4	0	19	0

Showing 1 to 9 of 9 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:

	Pardosa_ID	Measurement	Activity	duplicatedRow	originalRowID	value_Size	ngramID_Size
1	14	1	2	false	1	1.6	n8Seq1
2	14	2	11	false	2	1.6	n8Seq1
3	14	3	3	false	3	1.6	n8Seq1
4	14	4	8	false	4	1.6	n8Seq1
5	14	5	7	false	5	1.6	n8Seq1
6	14	6	8	false	6	1.6	n8Seq1
7	14	7	15	false	7	1.6	n8Seq1
8	14	8	13	false	8	1.6	n8Seq1
9	8	1	6	false	9	1.2	n8Seq6
10	8	2	2	false	10	1.2	n8Seq6
11	8	3	7	false	11	1.2	n8Seq6
12	8	4	8	false	12	1.2	n8Seq6
13	8	5	16	false	13	1.2	n8Seq6
14	8	6	21	false	14	1.2	n8Seq6
15	8	7	6	false	15	1.2	n8Seq6
16	8	8	8	false	16	1.2	n8Seq6
17	19	1	6	false	17	1.6	n15Seq4
18	19	2	10	false	18	1.6	n15Seq4
19	19	3	11	false	19	1.6	n15Seq4
20	19	4	7	false	20	1.6	n15Seq4
21	19	5	3	false	21	1.6	n15Seq4
22	19	6	8	false	22	1.6	n15Seq4
23	19	7	5	false	23	1.6	n15Seq4
24	19	8	17	false	24	1.6	n15Seq4
25	7	1	8	false	25	1.6	n15Seq4
26	7	2	21	false	26	1.6	n15Seq4
27	7	3	24	false	27	1.6	n15Seq4
28	7	4	25	false	28	1.6	n15Seq4
29	7	5	13	false	29	1.6	n15Seq4
30	7	6	9	false	30	1.6	n15Seq4

	Pardosa_ID	Measurement	Activity	duplicatedRow	originalRowID	value_Size	ngramID_Size
31	7	7	15	false	31	1.6	n15Seq4
32	7	8	8	false	32	1.6	n15Seq4
33	9	1	12	false	33	1.4	n16Seq1
34	9	2	14	false	34	1.4	n16Seq1
35	9	3	6	false	35	1.4	n16Seq1
36	9	4	7	false	36	1.4	n16Seq1
37	9	5	22	false	37	1.4	n16Seq1
38	9	6	8	false	38	1.4	n16Seq1
39	9	7	3	false	39	1.4	n16Seq1
40	9	8	11	false	40	1.4	n16Seq1
41	10	1	13	false	41	1.3	n16Seq1
42	10	2	16	false	42	1.3	n16Seq1
43	10	3	17	false	43	1.3	n16Seq1
44	10	4	7	false	44	1.3	n16Seq1
45	10	5	5	false	45	1.3	n16Seq1
46	10	6	20	false	46	1.3	n16Seq1
47	10	7	19	false	47	1.3	n16Seq1
48	10	8	15	false	48	1.3	n16Seq1
49	12	1	13	false	49	1.1	n7Seq10
50	12	2	13	false	50	1.1	n7Seq10
51	12	3	12	false	51	1.1	n7Seq10
52	12	4	17	false	52	1.1	n7Seq10
53	12	5	18	false	53	1.1	n7Seq10
54	12	6	6	false	54	1.1	n7Seq10
55	12	7	8	false	55	1.1	n7Seq10
56	12	8	11	false	56	1.1	n7Seq10
57	1	1	14	false	57	1.4	n16Seq2
58	1	2	20	false	58	1.4	n16Seq2
59	1	3	21	false	59	1.4	n16Seq2
60	1	4	16	false	60	1.4	n16Seq2
61	1	5	20	false	61	1.4	n16Seq2
62	1	6	23	false	62	1.4	n16Seq2
63	1	7	17	false	63	1.4	n16Seq2

	Pardosa_ID	Measurement	Activity	duplicatedRow	originalRowID	value_Size	ngramID_Size
64	1	8	18	false	64	1.4	n16Seq2
65	11	1	14	false	65	1.8	n16Seq2
66	11	2	12	false	66	1.8	n16Seq2
67	11	3	17	false	67	1.8	n16Seq2
68	11	4	18	false	68	1.8	n16Seq2
69	11	5	16	false	69	1.8	n16Seq2
70	11	6	17	false	70	1.8	n16Seq2
71	11	7	18	false	71	1.8	n16Seq2
72	11	8	11	false	72	1.8	n16Seq2
73	3	1	15	false	73	1.5	n7Seq11
74	3	2	11	false	74	1.5	n7Seq11
75	3	3	15	false	75	1.5	n7Seq11
76	3	4	16	false	76	1.5	n7Seq11
77	3	5	12	false	77	1.5	n7Seq11
78	3	6	19	false	78	1.5	n7Seq11
79	3	7	18	false	79	1.5	n7Seq11
80	3	8	11	false	80	1.5	n7Seq11
81	15	1	15	false	81	1.4	n16Seq1
82	15	2	16	false	82	1.4	n16Seq1
83	15	3	17	false	83	1.4	n16Seq1
84	15	4	12	false	84	1.4	n16Seq1
85	15	5	18	false	85	1.4	n16Seq1
86	15	6	13	false	86	1.4	n16Seq1
87	15	7	11	false	87	1.4	n16Seq1
88	15	8	5	false	88	1.4	n16Seq1
89	4	1	17	false	89	1.3	n16Seq1
90	4	2	4	false	90	1.3	n16Seq1
91	4	3	9	false	91	1.3	n16Seq1
92	4	4	14	false	92	1.3	n16Seq1
93	4	5	5	false	93	1.3	n16Seq1
94	4	6	23	false	94	1.3	n16Seq1
95	4	7	24	false	95	1.3	n16Seq1
96	4	8	17	false	96	1.3	n16Seq1

	Pardosa_ID	Measurement	Activity	duplicatedRow	originalRowID	value_Size	ngramID_Size
97	13	1	17	false	97	1.8	n8Seq4
98	13	2	15	false	98	1.8	n8Seq4
99	13	3	16	false	99	1.8	n8Seq4
100	13	4	13	false	100	1.8	n8Seq4

Showing 1 to 100 of 152 entries

Previous 1 2 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:

Pardosa_ID

run random reordering

randomisation result:

number of datapoints that are part of a duplicate sequence

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