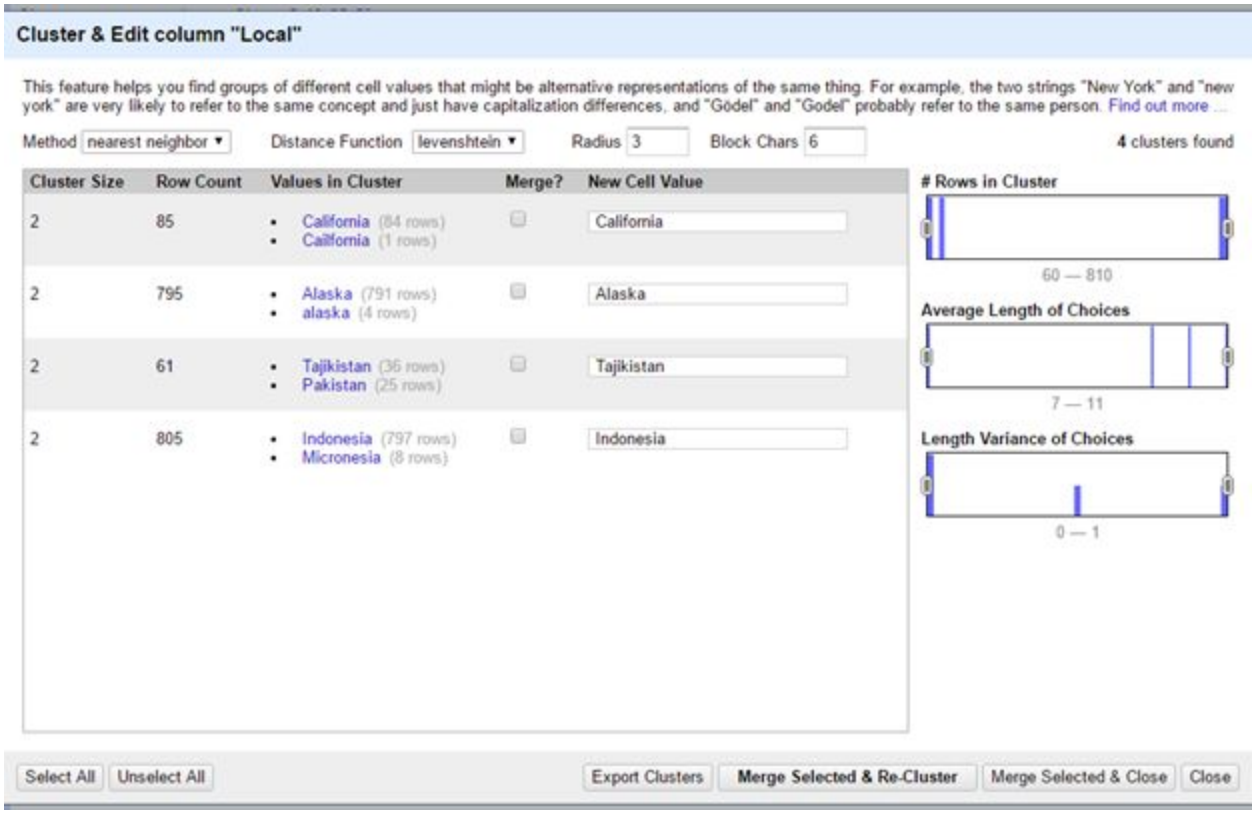
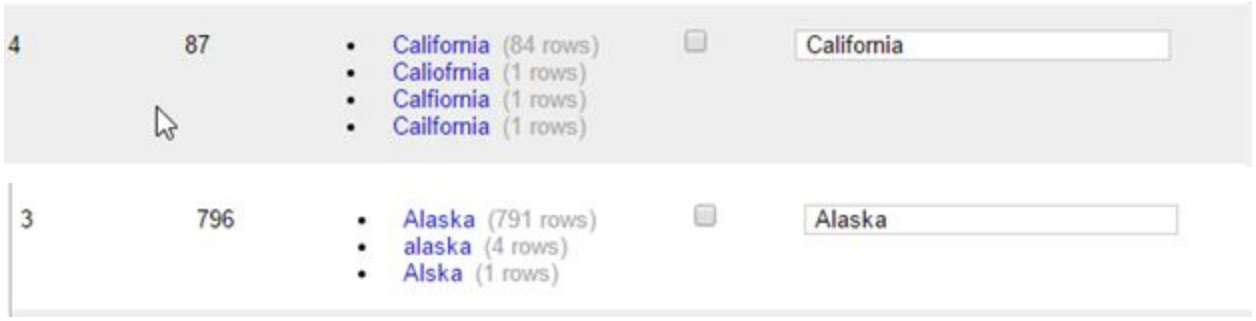
Adam Reilly
Lab 8

1. 5,377 rows do not have states (This is the number of "blank" states listed in the text facet).
2. 4,362 rows do not have zip codes
3. There are 380,136 valid zip codes and 4,362 blank zip codes (which I would call invalid)
4. This is what occurs when radius is changed to 3.

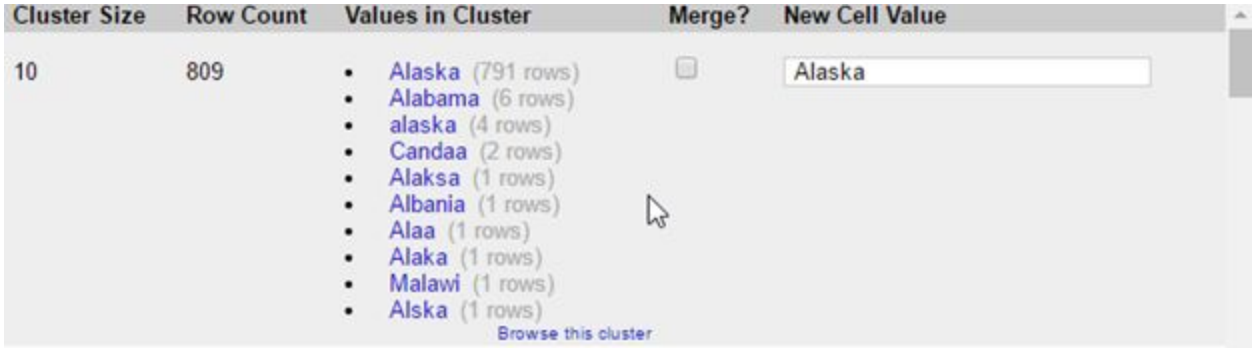


I would merge the first and second clusters, but not the third or fourth.

5. Changing the block characters to 2 greatly expands the number of clusters. Two clusters that I would choose to change are:



6. When you cluster on a place column, OpenFile groups together words based on letter similarity and proximity. Block will determine the number of letters that need to match in words for them to be included. I believe this is because we opted for Levenhstein for “distance”, and the radius determines the maximum allowable distance. An additional functionality that could be provided is to pick and choose words to match when instead of making you pick a whole cluster. For example of the flaw, see below. You can only merge the entire cluster, not just the words that match.



7. Here is a schema of the Levenshtein’s distance for number 7.

		1	2	3	4	5	6	7	8	9	10
			G	U	M	B	A	R	R	E	L
1		0	1	2	3	4	5	6	7	8	9
2	G	1	0	1	2	3	4	5	6	7	8
3	U	2	1	0	1	2	3	4	5	6	7
4	N	3	2	1	1	2	3	4	5	6	7
5	B	4	3	2	2	1	2	3	4	5	6
6	A	5	4	3	3	2	1	2	3	4	5
7	R	6	5	4	4	3	2	1	2	3	4
8	E	7	6	5	5	4	3	2	2	2	3
9	L	8	7	6	6	5	4	3	3	3	2
10	L	9	8	7	7	6	5	4	4	4	3