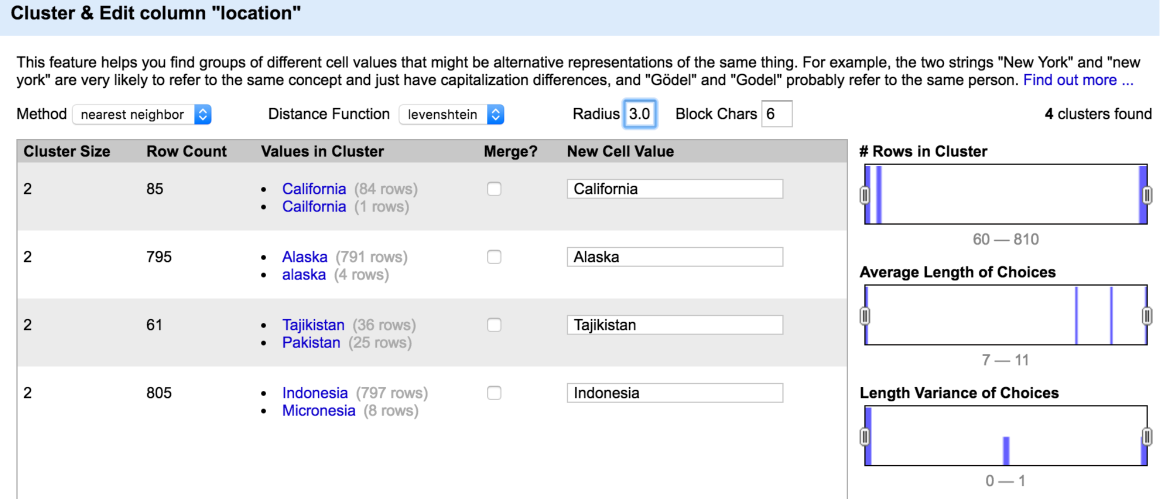
*1. How many rows are missing a value in the "State" column? Explain how you came up with the number.*
5377. Text Facet shows 5377 blank 'state' cells.

*2. How many rows with missing ZIP codes do you have?*
4362. Numeric Facet shows 4362 blank 'zip code' cells.

*3. If you consider all ZIP codes less than 99999 to be valid, how many valid and invalid ZIP codes do you have, respectively?*
Use this filter: if (value >99998, value, null) for numeric facet it shows 34961 invalid ('99999') values.
Use an equation: valid zip code number = 384498-4362-34961, the number of valid zip codes is 345175

*4. Change the radius to 3.0. What happens? Do you want to merge any of the resulting matches?*
I will not merge results because radius 3.0 tolerates more types of errors, trying to cluster unrelated words.



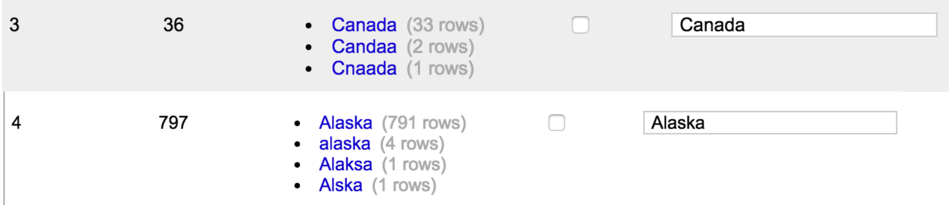
**Cluster & Edit column "location"**

This feature helps you find groups of different cell values that might be alternative representations of the same thing. For example, the two strings "New York" and "new york" are very likely to refer to the same concept and just have capitalization differences, and "Gödel" and "Godel" probably refer to the same person. Find out more ...

Method nearest neighbor   Distance Function levenshtein   Radius 3.0   Block Chars 6   **4** clusters found

| Cluster Size | Row Count | Values in Cluster | Merge? | New Cell Value |
|---|---|---|---|---|
| 2 | 85 | • California (84 rows)<br>• Cailfornia (1 rows) | ☐ | California |
| 2 | 795 | • Alaska (791 rows)<br>• alaska (4 rows) | ☐ | Alaska |
| 2 | 61 | • Tajikistan (36 rows)<br>• Pakistan (25 rows) | ☐ | Tajikistan |
| 2 | 805 | • Indonesia (797 rows)<br>• Micronesia (8 rows) | ☐ | Indonesia |

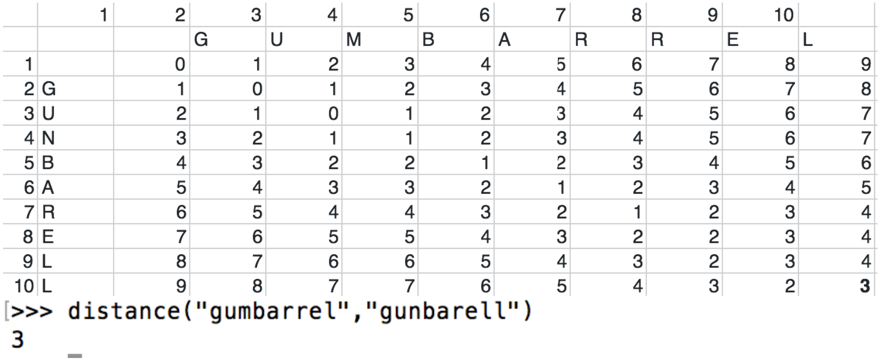
*5. Change the block size to 2. Give two examples of new clusters that may be worth merging.*
radius 2 block 2. Two examples are shown in two snapshots.



| | | | | |
|---|---|---|---|---|
| 3 | 36 | • Canada (33 rows)<br>• Candaa (2 rows)<br>• Cnaada (1 rows) | ☐ | Canada |
| 4 | 797 | • Alaska (791 rows)<br>• alaska (4 rows)<br>• Alaksa (1 rows)<br>• Alska (1 rows) | ☐ | Alaska |

*6. Explain in words what happens when you cluster the "place" column, and why you think that happened. What additional functionality could OpenRefine provide to possibly deal with the situation?*
It takes too long to finish grouping because too many different addresses while they have same or similar state number, giving rise to too many combinations.
We can create a location column and town column, then cluster and merge cells in those columns. Use the results to cluster the place column.

*7. Submit a representation of the resulting matrix from the Levenshtein edit distance calculation. The resulting value should be correct.*
*Results are shown in two snapshots. Python codes are provided, too.*



| | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | 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 | 3 | 4 |
| 9 | L | 8 | 7 | 6 | 6 | 5 | 4 | 3 | 2 | 3 | 4 |
| 10 | L | 9 | 8 | 7 | 7 | 6 | 5 | 4 | 3 | 2 | **3** |

```
>>> distance("gumbarrel","gunbarell")
3
```