USING OPENREFINE TO CLUSTER AND EDIT STRINGS

OpenRefine (previously also known as Google Refine) is a free, open-source software used for data cleaning as well as performing both basic and advanced cell transformations. This can be useful if you are working with large raw data files and need to systematically clean string or numeric variables (say, due to the absence of a unique identifier), or want a quick overview of the data before running further analysis.

OpenRefine can be a useful complementary tool to other statistical software, as part of the data cleaning process. While other softwares have a variety of string cleaning options, OpenRefine can be a more efficient alternative.

For example, in the simplest case, if we have multiple observations referring to Coder's Corner in different ways, such as 'Coders Corner', 'Coder's Korner', 'Corner Coder's', and 'Coder's Corner', multiple algorithms can be used in OpenRefine to group them into a cluster. These clusters can be used for subsequent analysis.

Getting started

The most recent version of OpenRefine (release 3.4.1 at the time of writing), can be downloaded at: https://openrefine.org/download.html. You need only choose the download 'kit' corresponding to your operating system and follow the installation prompts.

When working with large datasets, files close to 1GB or above, I recommend increasing the RAM allocated to OpenRefine to ensure the file is processed and uploaded smoothly. Windows users can do this by opening the *openrefine.exe* file and edit the following lines in the *openrefine.l4j.ini* file:

```
# max memory heap size

-Xmx1024M
```

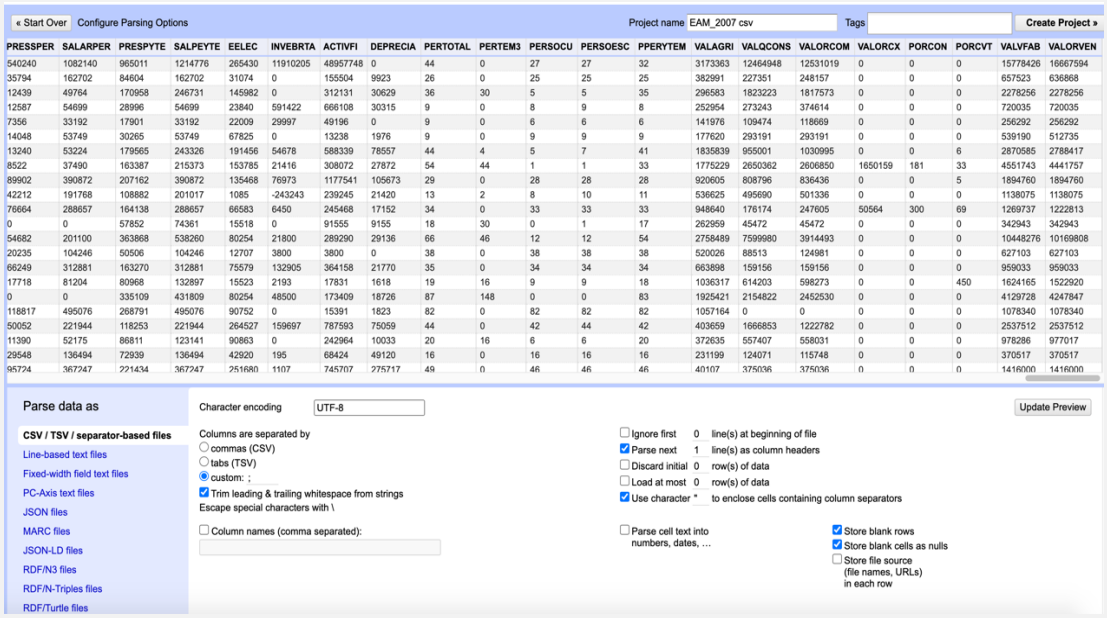
The last line specifies the memory allocated to OpenRefine in mebibytes (where 1 MiB ~= 1.049 MB). To increase memory, change '1024' to a larger number, e.g. to allocate 2GB of memory, update the last line to `-Xmx2048M`. You should also ensure your system is using the 64-bit version of Java.

If using a Mac, hold down the control key while clicking on the *OpenRefine.app* file, select *'show package contents'*, click on the *'Contents'* folder and open *'Info.plist'* with a text editor to change "`-Xmx2048M`" to the memory needed.

Importing the data

To get started, we import a dataset containing manufacturing firm data and firm locations. We can do this by creating a new project on the workspace.

The software can parse a large variety of file types, not limited to text-based files, with a full list of options listed in the bottom left panel.



PRESSPER	SALARPER	PRESPYTE	SALPEYTE	EELEC	INVEBRTA	ACTIVFI	DEPRECIA	PERTOTAL	PERTEM3	PERSOCU	PERSOESC	PPERYTEM	VALAGRI	VALQCONS	VALORCOM	VALORCX	PORCON	PORCVT	VALVFAB	VALORVEN
540240	1082140	965011	1214776	265430	11910205	48957748	0	44	0	27	27	32	3173363	12464948	12531019	0	0	0	15778426	16667594
35794	162702	84604	162702	31074	0	155504	9923	26	0	25	25	25	382991	227351	248157	0	0	0	657523	636868
12439	49764	170958	246731	145982	0	312131	30629	36	30	5	5	35	296583	1823223	1817573	0	0	0	2278256	2278256
12587	54699	28996	54699	23840	591422	666108	30315	9	0	8	9	8	252954	273243	374614	0	0	0	720035	720035
7356	33192	17901	33192	22009	29997	49196	0	9	0	6	6	6	141976	109474	118669	0	0	0	256292	256292
14048	53749	30265	53749	67825	0	13238	1976	9	0	9	9	9	177620	293191	293191	0	0	0	539190	512735
13240	53224	179565	243326	191456	54678	588339	78557	44	4	5	7	41	1835839	955001	1030995	0	0	6	2870585	2788417
8522	37490	163387	215373	153785	21416	308072	27872	54	44	1	1	33	1775229	2650362	2606850	1650159	181	33	4551743	4441757
89902	390872	207162	390872	135468	76973	1177541	105673	29	0	28	28	28	920605	808796	836436	0	0	5	1894760	1894760
42212	191768	108882	201017	1085	-243243	239245	21420	13	2	8	10	11	536625	495690	501336	0	0	0	1138075	1138075
76664	288657	164138	288657	66583	6450	245468	17152	34	0	33	33	33	948640	176174	247605	50564	300	69	1269737	1222813
0	0	57852	74361	15518	0	91555	9155	18	30	0	1	17	262959	45472	45472	0	0	0	342943	342943
54682	201100	363868	538260	80254	21800	289290	29136	66	46	12	12	54	2758489	7599980	3914493	0	0	0	10448276	10169808
20235	104246	50506	104246	12707	3800	3800	0	38	0	38	38	38	520026	88513	124981	0	0	0	627103	627103
66249	312881	163270	312881	75579	132905	364158	21770	35	0	34	34	34	663898	159156	159156	0	0	0	959033	959033
17718	81204	80968	132897	15523	2193	17831	1618	19	16	9	9	18	1036317	614203	598273	0	0	450	1624165	1522920
0	0	335109	431809	80254	48500	173409	18726	87	148	0	0	83	1925421	2154822	2452530	0	0	0	4129728	4247847
118817	495076	268791	495076	90752	0	15391	1823	82	0	82	82	82	1057164	0	0	0	0	0	1078340	1078340
50052	221944	118253	221944	264527	159697	787593	75059	44	0	42	44	42	403659	1666853	1222782	0	0	0	2537512	2537512
11390	52175	86811	123141	90863	0	242964	10033	20	16	6	6	20	372635	557407	558031	0	0	0	978286	977017
29548	136494	72939	136494	42920	195	68424	49120	16	0	16	16	16	231199	124071	115748	0	0	0	370517	370517
95724	367247	221434	367247	251680	1107	745707	275717	49	0	46	46	46	40107	375036	375036	0	0	0	1416000	1416000

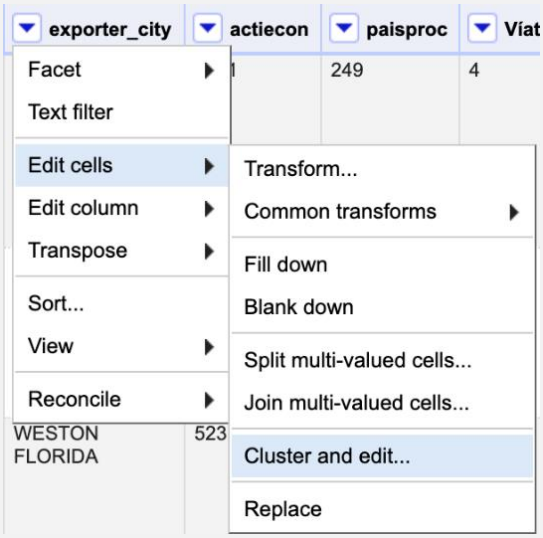
There are numerous options listed along the bottom of the project preview page that allows you to customise the data delimiter and choose whether to trim leading and trailing spaces. When satisfied with the options selected, you can go ahead and create your project. One handy feature of OpenRefine is that it has an undo/redo panel that appears on the left-hand side of the workspace once you've created your project that allows you to track changes and revert anything done to the dataset.

Cluster and Edit

OpenRefine's *'cluster and edit'* option is useful for cleaning messy string or numeric variables. You may find yourself working with large, raw data files that contain manually inputted string variables without an identifier for each unique entity. For example, you have firm-level data where a firm's name has multiple typos across different observations, random capitalizations, different word ordering, and random trailing or leading spaces, but nevertheless, multiple observations refer to the same entity.

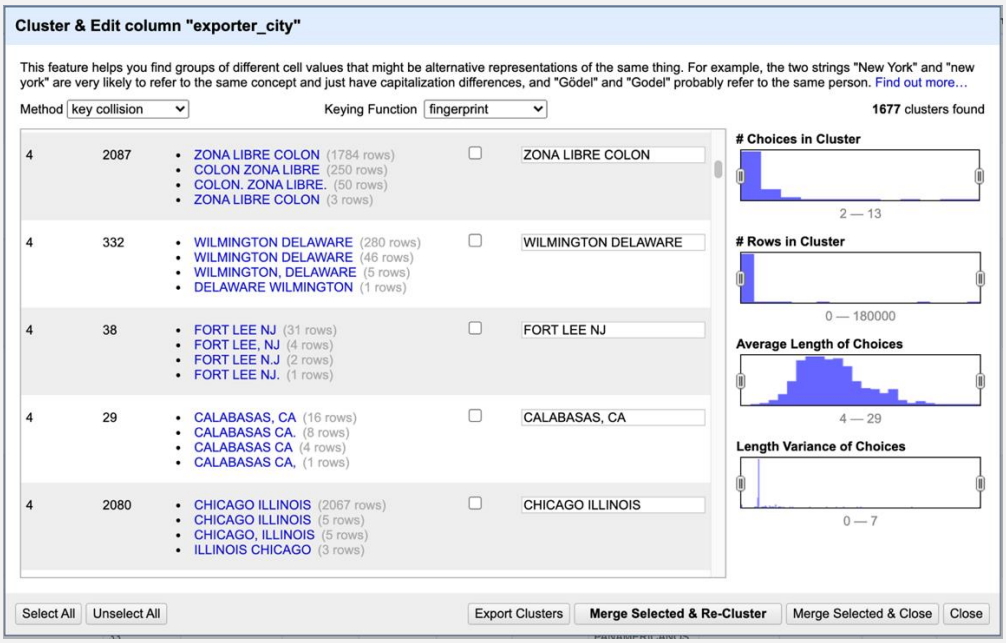
While Stata can perform both basic and advanced string cleaning functions, it falls short where there is no unique identifier for a string column of interest. Accounting for each error type across observations makes it difficult to systematically clean raw data consistently. Although leading or trailing spaces can quickly be mopped up in Stata, different word ordering and/or multiple typos across string observations that refer to the same entity cannot be handled easily in Stata. This makes it difficult to catch all the observations in the dataset that belong to the same entity.

OpenRefine has two solutions for messy strings. For large datasets, with too many unique values to be displayed via the *'facet'* option, which allows users to apply a particular value to all cells contained in a facet grouping, one can instead use the *'cluster and edit'* function. Clicking on the drop-down arrow of the column of interest, where here we'd like to clean the *'exporter_city'* variable, we select the *'cluster and edit'* option.

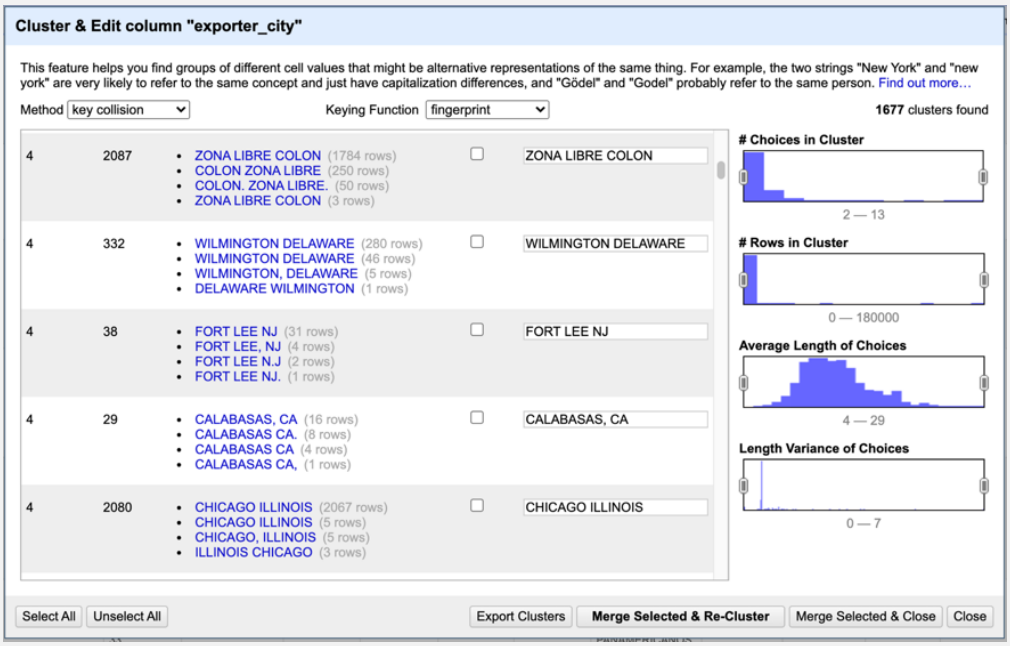


This takes you to the following page. Here city names have been grouped where, by default, OpenRefine applies the most stringent clustering method of key collision fingerprinting. Key collision operations are very fast when working with large datasets. Fingerprinting is the most stringent clustering method, and so the least likely to cluster observations that refer to different entities. It automatically applies more basic cleaning operations such as removing whitespaces, ignoring punctuation and uppercase/lowercase inconsistencies when clustering strings, removing character accents if working with data in foreign languages, and sorts words alphabetically. The last operation is useful in preventing clustering issues when typos include different rearrangements of the same (or similar) words.

Depending on the nature of the issue at hand there are alternative clustering methods available, including nearest neighbour matching, Levenshtein distance, and prediction by partial matching.



Users can edit the cluster name that OpenRefine will apply to each observation in a given cluster. The right-hand panel specifies the number of clusters found, as well as useful summary statistics. This provides additional flexibility in adjusting the observations included in a particular cluster along multiple dimensions such as the number of choices in the cluster, the rows in the cluster, the average length of choices etc.



For example, we can see that CHICAGO ILLINOIS and ILLINOIS CHICAGO are grouped into the same cluster. While it may seem simple enough to apply this in statistical packages in this scenario, it becomes exceedingly more complicated when the string column contains multiple words. In this case, a few misspellings or changes in the ordering of the firm's name (or the absence of certain words that are present in other observation, say), make identifying all observations that belong to the same entity tricky.

OpenRefine can also be used for a range of other purposes which can be found here: https://docs.openrefine.org/

Diana Beltekian, PhD Candidate in Economics, University of Nottingham
10 February 2021