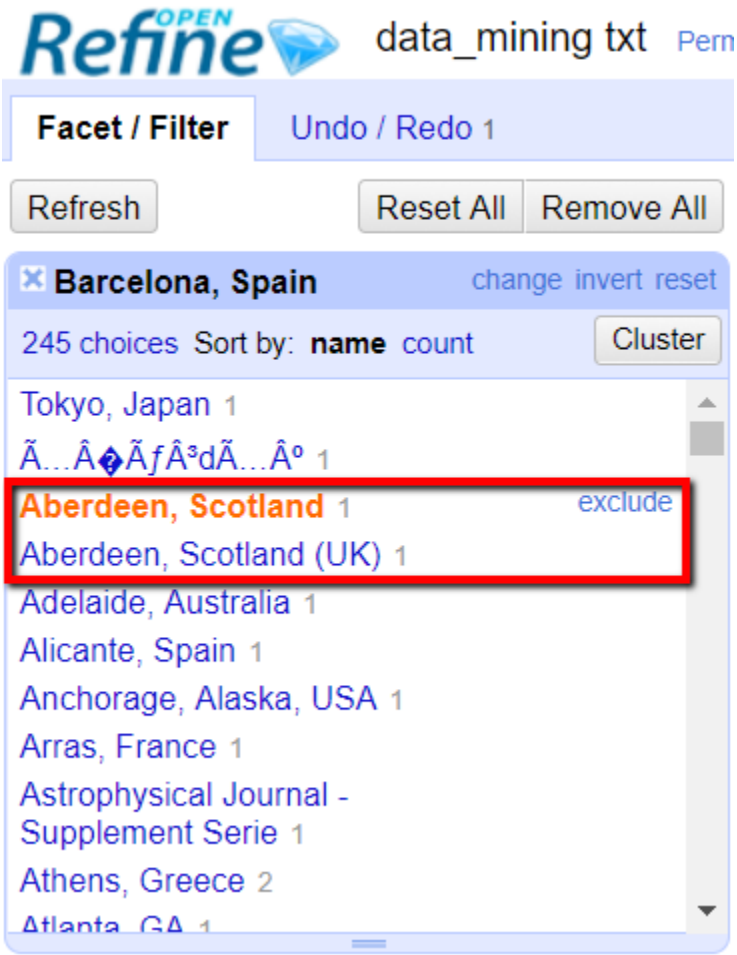
Casey Kongpanickul
Assignment Phase 1- Report

The strategy used to clean the data was first to get an idea of what it looked like overall. After looking at a few YouTube videos on how to use Open Refine it was clear that the best way to accomplish this quickly would be to use a text facet and quickly browse through the data of each column.
My strategy for cleaning the data is to go column by column. Things I want to take care of are duplicates, error values, and consolidating data that is unnecessarily separated. I chose to keep the N/A values because I did not wait to lose that data just yet. My thought processes is that I don't want to get rid of any data that I may want later and if I really don't need it later I can discard it then.

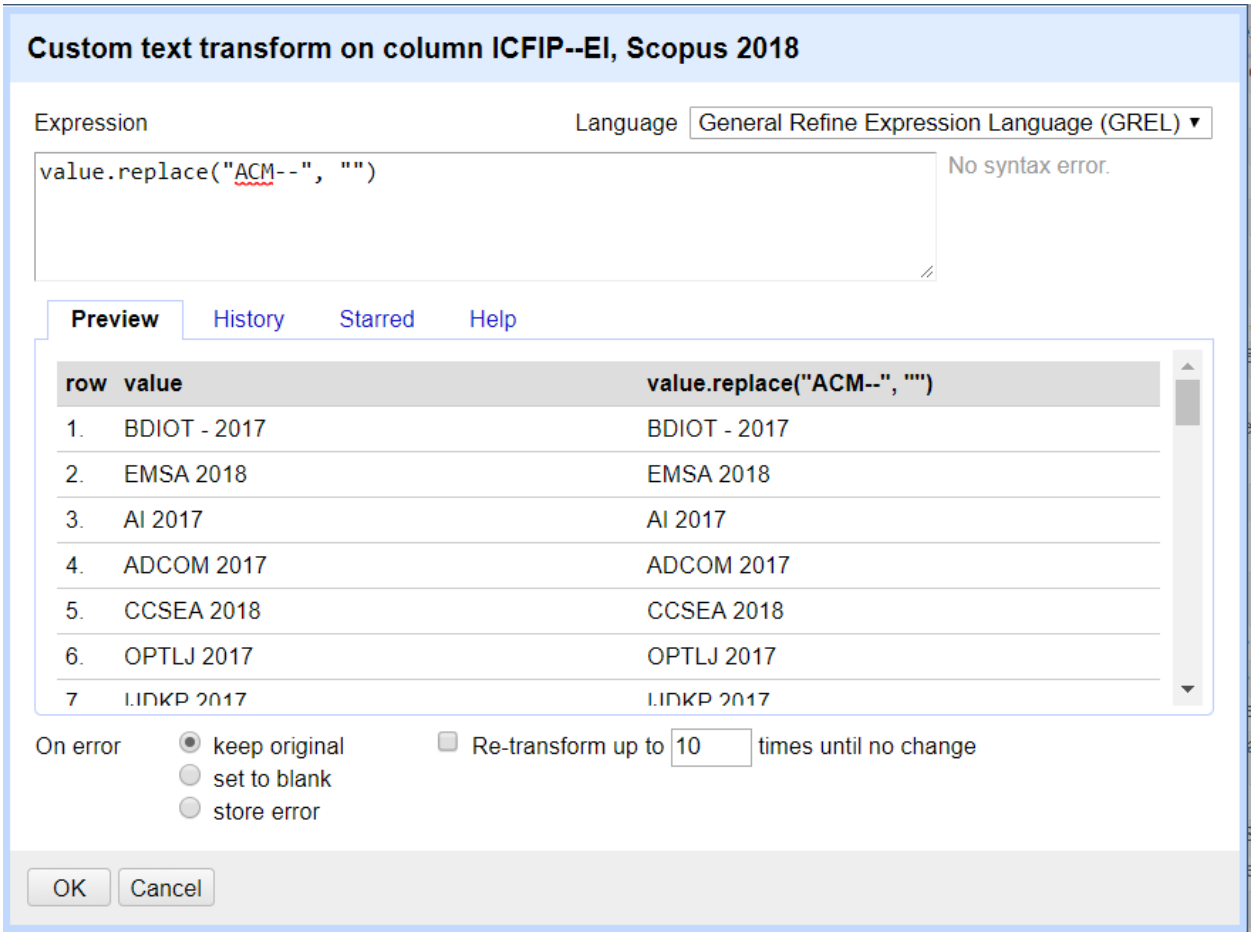
The location column would be the most straight forward to clean up since the values contained in the column would be the most familiar and easiest for spotting errors and inefficiencies. This data was for the most part well formatted but a name of a single place could be typed out different ways causing more rows than necessary. Using the text facet, I quickly sifted through the data to edit certain columns to conform to others.



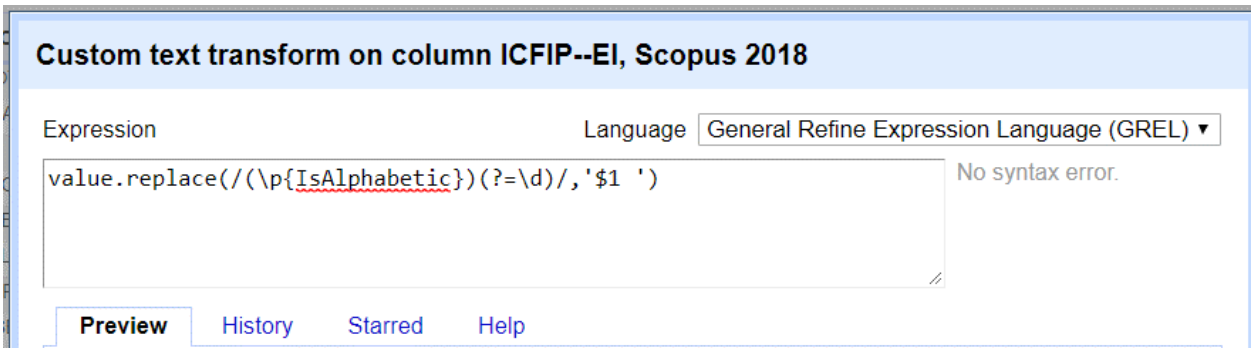
Since this was my first time using the program I was curious to see how quickly I could go through this data and group up the necessary rows. The time it took was 10 minutes which was a little tedious but a rather quick way to group up the rows with the location. I had dropped down to 184 rows.

The second column to be cleaned was the acronym column because immediately I noticed extra information that could be easily removed. Many rows had acronym names but also a prefix including ACM. To clear these up, the transform function was used (value.replace(arguments))

This transformation is essential the same as doing a find and replace all, like in Excel. There were a few instances that I used this approach to scrub the data in this column as well as the name column. For example, many of the values had the unnecessary information of the year which needed to be erased.

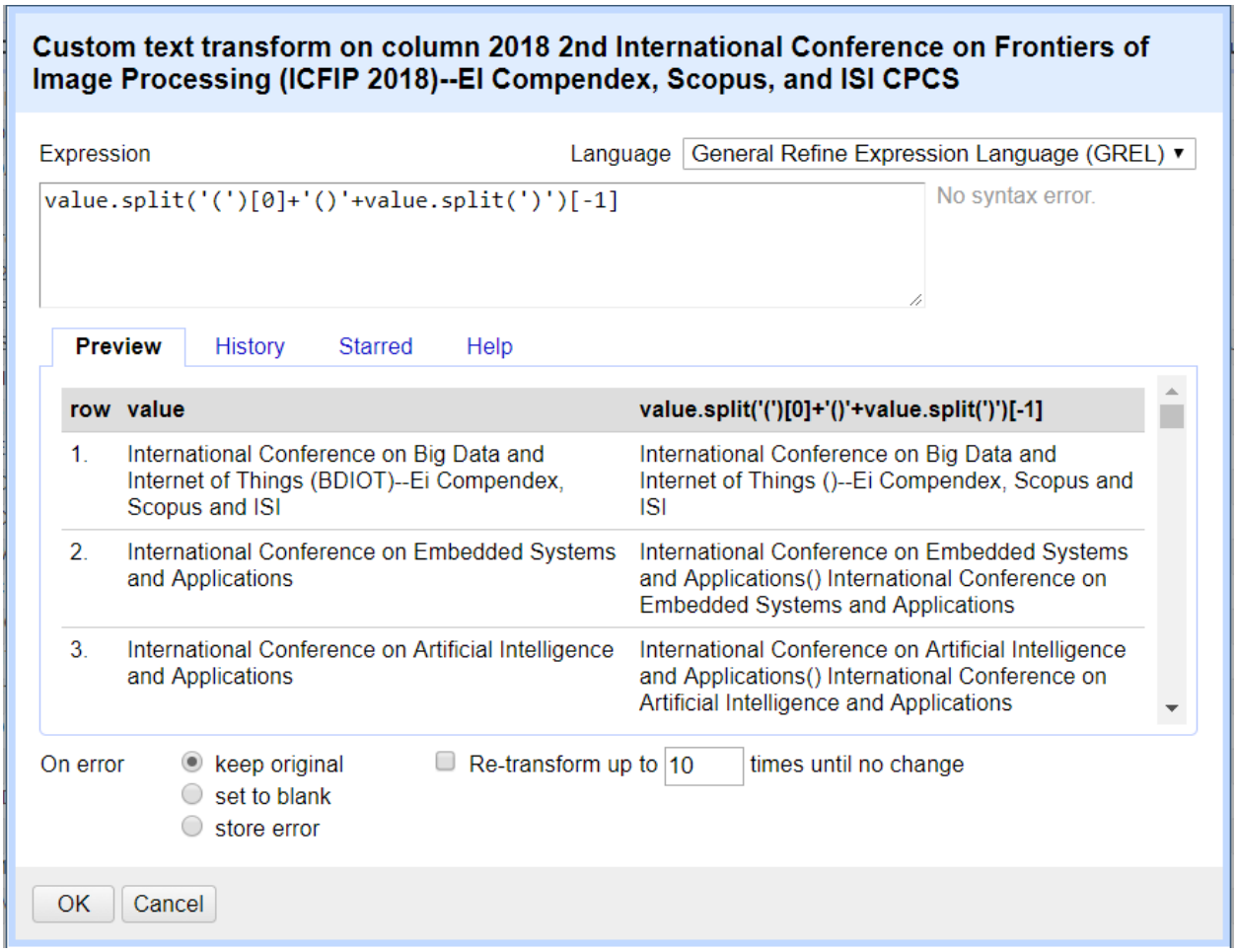


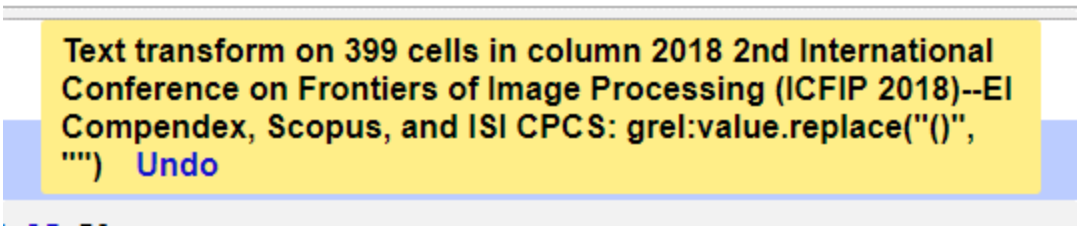
It was clear that some of the formatting in the values could be improved. Some of the cells had numbers and values beside one another with no space in between. Again, the transform function was used using the General Refine Expression Language. This time using regular expressions to match to letters and then create a space to the numbers.



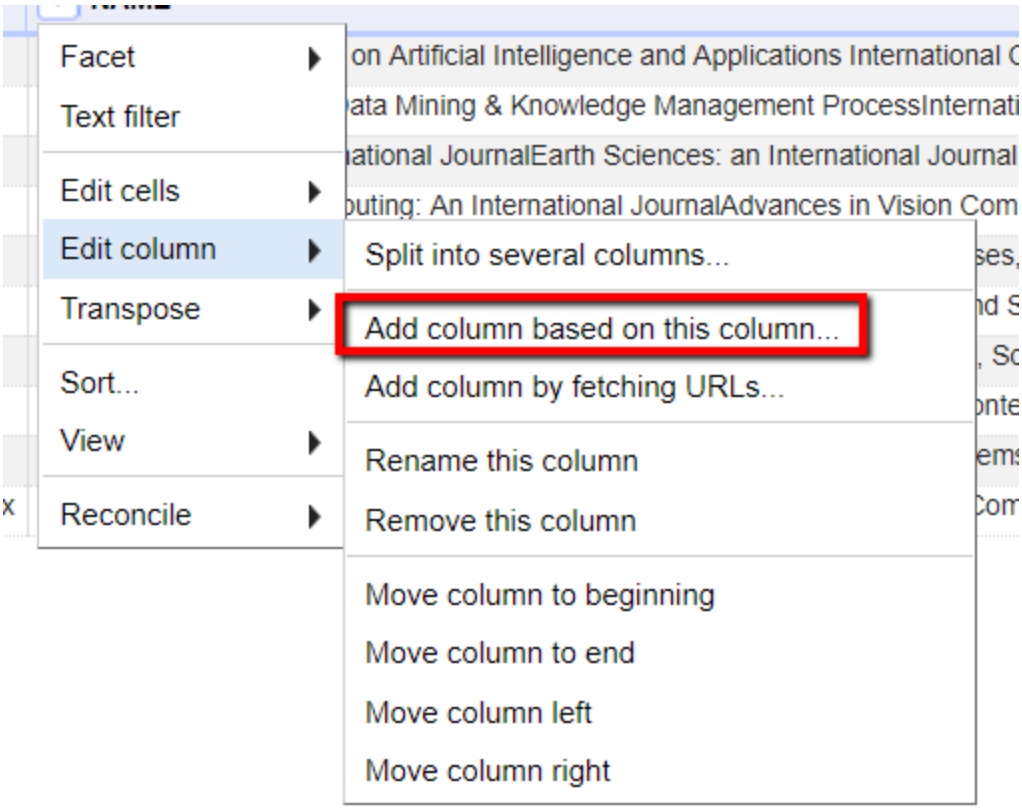
Another instance of needing to scrub acronyms from the data.

Getting a little bit more complex and using regular expressions it was clear that values between parentheses were not needed. This was taken care of in two steps. The first step was to clean the data of the values in between parentheses followed by simply removing the parentheses that were left over. I struggled with writing the regular expression myself for a while until finally turning to the internet to help find a function that would accomplish what I needed to do as seen below.

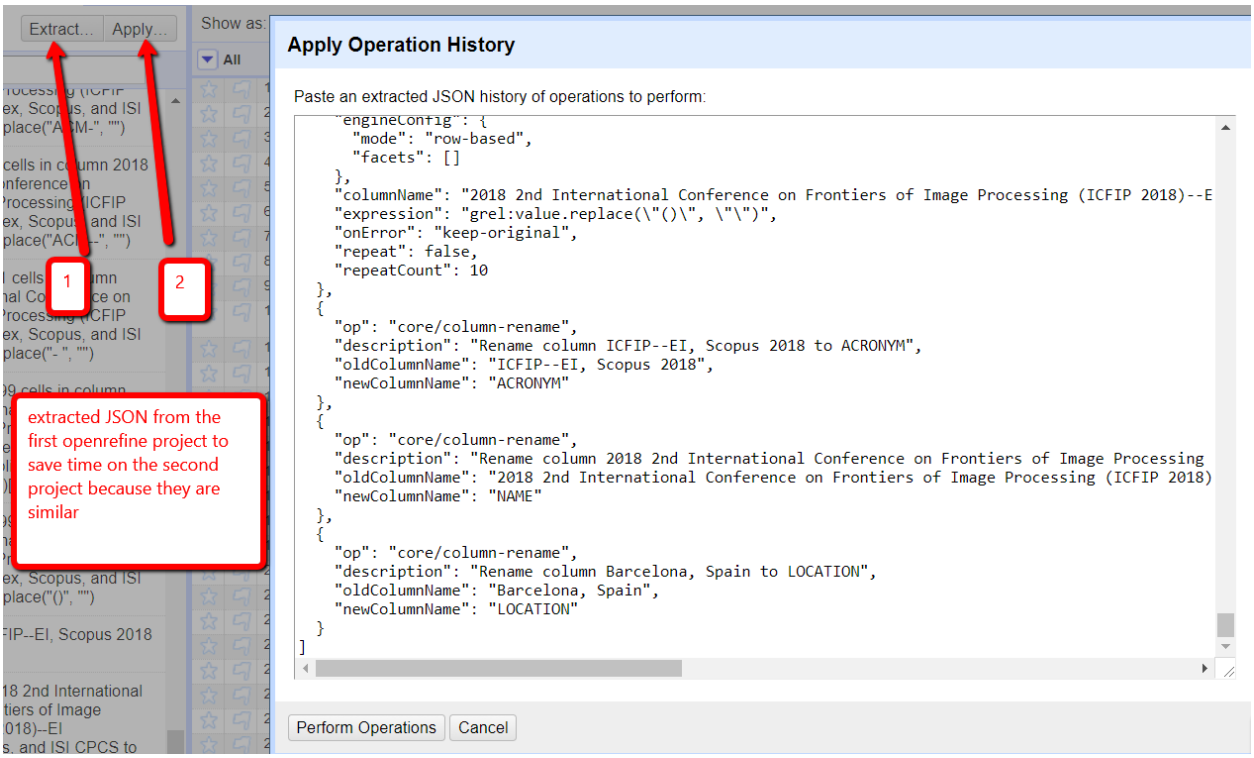




It was also to get rid of the .Journals since we are only interested in conferences. To accomplish this the functionality to create a new column bases on a column was used. The function was a .contains to check if the column contains the value I was looking for. From there it is possible to filter for only the value you needed and then delete the rest.



My crawler was run 4 times to crawl all the pages needed. This means that I have 4 very similar data sets. I used the functionality from Open Refine to extract the JSON code of instructions to clean the data set and applied it to my other outputs of the crawler. This helped but a large start on the data cleansing process. However to be thorough I followed a similar process manually on all subsequent datasets to make sure nothing was missed.



Ultimately, the hardest part of the phase 1 of this assignment was the crawling of the web page. I had a very difficult time extracting the data I needed from the html table because the table itself was not built in a very intuitive way. The rows were not specified as headers when necessary and the values were not stored by class name so it took a lot of iterations through each row to extract each piece of data needed. I assume this was the point of the exercise because data in the wild will never be perfect. Also, I am quite confident there is an easier way to clean the data using Open Refine and there is probably some advanced functionality that I may have missed out on.