**LEARNING JOURNAL WEEK 7-8**

**OPEN REFINE FOR SOCIAL SCIENCES.**

Wednesday (October 2)

**Objective:** To complete data carpentry for Open Refine.

**Action:** to complete the exercises by following the instructions provided on Data Carpentry Site.

**Set Up:**

As I couldn’t set up the openrefine in class, I took help of Sheridon to download, which took whole lot of time. Openrefine is downloaded from [www.openrefine.org](http://www.openrefine.org)

It requires java plug in. the tool started in a window and in order to work on the data the working tool is opened in a browser window.

**LESSON 1: INTRODUCTION**.

Ques: What is Open Refine useful for?

Ans: It is a powerful, free and open source tool that can be used for data cleaning. It will automatically track any steps allowing you to backtrack as needed and providing a record of all work done.

**LESSON 2: WORKING WITH OPEN REFINE**

Ques: How can we bring our data into Open Refine?

Ans: Data can be brought into Open Refine by importing a readable file, such as a .csv. It could be done using other file formats such as .tsv, .xls, .xlsx, and .xml.

Ques: How can we sort and summarize our Data?

Ans: Open Refine has Facet tool, using that we can view our data and then filter it down to what we want to Refine.

Ques: How can we find and correct errors in our raw data?

The program can identify identical data and inconsistencies by using the facet tool, hence the mistakes can be located.

**FACETS:**

Facets are an important tool for working in openrefine. Using a facet it is observed that the data in “village” column has discrepancies. These discrepancies can be corrected using the ‘edit’ in the facet.

**NUMERIC DATA IN OPENREFINE:**

It is important to note that, we have to transform data type from text to numeric when we have to work with data numerically.

**CLUSTERING:**

To correct the discrepancies in ‘villages’ column, the clustering function can be used in the facet.

Click ‘cluster’, window showing various methods of clustering data points opens with the outcomes.

In order to refine the data, we can choose any method and can do this process as many times until the data is refined.

But I faced a problem. In the ‘villages’ column, there is a field saying ‘49’, I was unable to merge it with any other name, so further information is required.

Again, there was a problem, ‘Chirdozo’ wasn’t merged as ‘Chirodzo’, so it had to be done manually.

**TRANSFORMING DATA:**

In the column ‘items owned’, there are certain irrelevant characters like that have to be removed in order to remove difficulties in faceting the columns.

‘transform’ function is used, here a GREL expression is used to execute a change.

In order to change and clean the column, the following GREL is used:

value.replace(”[”, ””)

value.replace(”]”, ””)

value.replace(”’”, ””)

value.replace(” ”, ””)

With multiple actions, these are combined with one string, that is :value.replace(”’”,””).replace(”[”,””).replace(”]”,””).replace(” ”,””).

Now the information is separated in the column only by ;

To find two items that are most commonly owned and least commonly owned:

Action: Select Sort by: count

Result: the most commonly owned items are phones and radio and least commonly owned are cars and computers.

Applying the same process to months lack food column, it is found that November is the most common month for the respondents to lack food.

The same cleanup can be applied to other columns, by using the previous GREL expressions in the transform window in the history tab.

**UNDO AND REDO**

We can undo or redo multiple steps in OpenRefine. It is a kind of record of all actions that we have applied while doing the project.

We can undo multiple steps and it can be checked before confirming and then continuing. Afterwards we can redo in the same way.

**Trim Leading and Trailing Whitespace**

OpenRefine provides a tool to remove blank characters. Using the common transform function to trim leading and trailing whitespace, Like when we facet the wall type column, a field is separated because some of the entries contain an extra space which is not noted by human eye, can be resolved.

**LESSON 3: FILTERING AND SORTING WITH OPEN REFINE**

**FILTER:** Filtering is a useful tool. The text filter is automatically applied to the facet and vice versa.

We can filter the data by applying a text facet on the respondent\_roof\_type column.

Now there are two data results that match the filter while filtering by ‘mabat’. The number of rows selected is highlighted.

**SORT:** The rows can be sorted from the dropdown menu from a column heading.

Apply sort to the column ‘gps: altitude’.

Select Sort…>numbers and select ‘smallest first’. It reveals few cells having 0 values, the reason might be the information not being recorded.

**SORTING BY MULTIPLE COLUMNS:**

Sort function can be applied to multiple columns.

Using this to find the correct village for the entry ‘49’ in the village column. Now edit is used to move the village column to the end, the village data can be compared to gps information. It could be implied that entry ‘49’ could be ‘Chirodzo’ but as other villages also have the similar gps coordinates it cannot be definitely concluded. But by comparing it to the interview date , it is seen that Chirodzo is the only village interviewed on the same date as the entry’49’, it is confirmed, so the change can be applied.

**LESSON 4: EXAMINING NUMBERS IN OPEN REFINE**

Transform function is used to change the way data has to be presented, i.e. from text to number.

NUMERIC FACET:

We can run a numeric facet to ‘years farm’ column after changing it to number. Here the information is presented in a numeric range and can be filtered through the adjustment in highlighted range.

Further, some check boxes appear at the bottom of the facet if a text cell is included in a numeric sort. By checking and unchecking the boxes, it changes the rows that appear on the screen.

**LESSON 5: USING SCRIPTS:**

OpenRefine work can be saved as a script.

In the undo/redo tab, this can be extracted by clicking Extract…

Then we can see a body of text that can be copied and saved.

The steps can be selected (to be applied to other datasets) by clicking the check boxes.

The list of operations can be applied to a new project.

Click apply button in the undo/redo tab, then copy and paste contents of our already created file, then the perform operations button will apply the work to the current project.

**LESSON 6: EXPORTING AND SAVING DATA FROM OPEN REFINE**

The next step is exporting the refined data.

**SAVE**

We can save our work continuously in OpenRefine and then we can resume our work.

**EXPORTING**

OBJECTIVE: export the project from OpenRefine.

ACTION:

* Click the export button in the top right
* Select export project

There was a continuous error coming. So I was not able to complete the step.

**LESSON 7: OTHER RESOURCES IN OPEN REFINE**

Open refine is a very useful tool. There are different online resources offered through www.openrefine .org. in addition there are other useful resources having useful information related to working with data.

**TASK 2**

**03/10/19 (7:30 pm)**

Proof of Concept Design.

As I couldn’t attend the week 7 class, everything was a mess. I didn’t have any idea what to do. After consulting Shawn I got a fair idea of the POC Design. Following his instructions and in the homework on Cloudstor I created 5 user stories, the prerequisites and the acceptance criteria to the best of my ability. I have uploaded it on Cloudstor and committed to github.