Module 3 Lab:
Data Wrangling with OpenRefine

[10pt] OpenRefine

a. Watch the videos on the OpenRefine's homepage for an overview of its features. Download and install the latest version of OpenRefine.

b. Import Dataset:
   1. Launch OpenRefine. It opens in a browser (127.0.0.1:3333).
   2. We use a products dataset from Mercari, derived from a competition on Kaggle (Mercari Price Suggestion Challenge). If you are interested in the details, please refer to the data description page. We have sampled a subset of the dataset as the given "properties.csv".
   3. Choose "Create Project" -> This Computer -> "properties.csv". Click "Next".
   4. You will now see a preview of the dataset. Click "Create Project" in the upper right corner.

c. Clean/Refine the data:
   Note: OpenRefine maintains a log of all changes. You can undo changes. See the "Undo/Redo" button on the upper left corner.

   i.a [1 pt] Select the "category_name" column and choose 'Facet by Blank' (Facet -> Customized Facets -> Facet by blank) to filter out the records that have blank values in this column. Provide the number of rows that return True. Remove these rows.

   i.b [1 pt] Split the column "category_name" into multiple columns without removing the original column. For example, a row with "Kids/Toys/Dolls & Accessories" in the category_name column, would be split across the newly created columns as "Kids", "Toys" and "Dolls & Accessories". Use the existing functionality in OpenRefine that creates multiple columns from an existing column based on a separator (i.e. in this case '/'). Provide the number of columns that are created in this operation. Remove the newly created columns that do not have values in all rows.

   ii. [2 pt] Select the column "name" and apply the Text Facet (Facet -> Text Facet). Click on the Cluster button which opens a window where you can choose different "methods" and "keying functions" to use while clustering. Choose the keying function that produces the highest number of clusters under the "Key Collision" method. Provide the number of

clusters found using this keying function. Click on 'Select All' and 'Merge Selected & Close'.

iii. [2 pt] Replace the null values in the "brand_name column" with the text "Unbranded" (Edit Cells -> Transform). Provide the General Refine Evaluation Language (GREL) expression used.

iv. [2 pt] Create a new column "high_priced" with the values 0 or 1 based on the "price" column with the following conditions: If the price is greater than 100, "high_priced" should be set as 1, else 0. Provide the GREL expression used to perform this.

v. [2 pt] Create a new column "has_offer" with the values 0 or 1 based on the "item_description" column with the following conditions: If it contains the text "discount" or "offer" or "sale", then set the value in "has_offer" as 1, else 0. Provide the GREL expression used to perform this.

Deliverables

- **properties_clean.csv** : Export the final table as a comma-separated values (.csv) file.
- **changes.json** : Submit a list of changes made to file in json format. Use the "*Extract Operation History*" option under the Undo/Redo tab to create this file.
- **Observations.txt** : A text file with answers to parts c.i.a, c.i.b, c.ii, c.iii, c.iv and c.v. Provide each answer in a new line.