

**Group Members:**

Muhammad Moiz Arif 15K-2146

Abdul Munim Khan 15K-2897

**Section:** GR-3

**Subject:** Data Science

***Assignment 1***

**Question 1:**

**Human-Centric Data Cleaning**

**Summery:**

In this research paper we understand that, if we looking at data cleaning techniques, we make the following observations: Human involvement in algorithm-driven: Human are the ultimate authority in verifying the accuracy of the data. Typically, humans are tightly coupled to the cleaning logic, i.e. humans are involved in ways that are dictated by the cleaning algorithm being used. It just as important in suggesting potential data error for humans to verify, it is important to make it easy for humans to identify faulty factors that have been involved to compute the inaccurate repairs. Secondly, Humans are not always right: Many human-driven data cleaning techniques assume that humans are perfect. There are several fine distinctions in the human feedback that are need to be analysis to effectively involve humans in the cleaning process. Assume that a human and an automatic agent participate to detect and repair the data. The question is that, how to effectively involve humans in the cleaning process? Four steps are: detection, repairing, validation, specification. Thirdly, Human cost Optimization: The system should be able to reason about the expertise of different humans when assigning cleaning tasks. While several research efforts involve the human in specific cleaning problem, there is no specified proposal that involves humans for general data cleaning. Fourthly, characterizing human expertise in cleaning tasks: Human interact in various ways in the cleaning process. The human budget for a data cleaning job j could be expressed as a combination of many factors including the maximum number of humans available to perform a certain task, the total money cost to spend to perform a task, etc. A human in a data cleaning scenario is represented as h: Role, Data, Cost, and Enterprise where Role is a role of the human, Data is a set of cells h is knowledgeable about, Cost is a cost of involving h, Enterprise is a score that reflects how good h is for the role Role in cells Data. You can understand how important is human. Now, we talk about task assignment, given the data cleaning job j for cells C, a pool of humans, say H, and a budget, say B, the framework should assign automatically cells in C to human in H. The assignment should guarantee the following properties: Coverage: if the job is to be performed by human only, every cell in c should be covered at least one human; Maximize Expertise: The assigned human to given cells should have good knowledge about these cells; Minimize cost: The human cost should not exceed Budget B. Now, we talk about cost optimization, minimizing the human cost to repair data has been the corner stone of numerous research efforts. We need to answer the following questions: When we have human and automatic cleaning agent, what are the consequences of involving one over the other on human cost and data quality? Given multiple humans that are assigned the same set of cells to repair, which human do we choose? How do we schedule different cleaning jobs in order to achieve an optimal human cost and data quality? Two ways of cost optimization, quantitative and qualitative. Quantitative cost optimization: It prefers automatic agent over humans. While in qualitative cost optimization: The strategy gives preferences to human over automatic tools. In conclusion, there are still many other questions that were not addressed in the vision such as data privacy, i.e. how can humans clean the data in the presence of privacy constraints? But we believe the proposed vision raises the central questions that we need to answer first to realize a human-centric data cleaning system.

**Question 2:**

**Detecting Data Errors:**

**Where are we and what needs to be done?**

**By:** Ziawasch Abedjan

We have tried two softwares, first was openRefine, but we was not succeeded to accomplish goal, so we have moved to trifacta(online). We have created online free account. We simply import data1.csv, data2.csv and data3.csv of lab 3. After import we have created flow and add all datasets to flow. One by one add recipe and edit it in data1.csv we have taken average of column B and fill missing cell. In data2.csv we have replaced delimited value with 1(it could be any value). In data3.csv we won’t find any error.

**Comparison between Trifacta and Python:**

Trifacta has better UI than simple Python console, and given presets of option reduce coding, save time and human errors. Data in Python is tedious and time consuming. It meant continually checking for inconsistencies and then scripting the remaining variations. Once we adopted Trifacta Wrangler, it took less than a minute to clean the entire dataset.

**Benefits of Trifacta:**

**Visual Benefit:**

Missing values are highlight in red color and all changes are done on run time, option of undo and redo has given to prevent extra efforts. Major options are given in top panel when edit recipe. Grid and column views help in understanding of what you have in first look.

**Time Consumption:**

Less code, simply leads to less time consumption. Direct manipulation of what is needed.

**Less Code:**

To find missing values or find delimited values etc. giving command helps you, which make work easy.

**Easy to Export:**

Given export types like csv and JSON helps in transforming data.

**Structuring:**

Trifacta helps in assigning proper values and finding anomalies. Like we have seen in data1.csv and data2.csv of lab 3

**Cleaning:**

To make data model ready it gives text standardization, replace and merge etc.

**Disadvantage of Trifacta:**

As we have very little amount of data, so we can not judge very much about disadvantages of trifacta, but some we have found are python is a language and having command line interface work more efficiently than trifacta. You won’t have restricted with given parameters, you do what you want.

**Results**:

We have attached data1.csv, data2.csv and data3.csv for reference. These files have exported from trifacta website.