CS481 Data Science

Assignment 1

February 16, 2020

Bilal Hyder 16I0262
Bilal Khan 16K3778

Q1 Human Centric Data Cleansing Summary

The document/research paper discusses about Data Cleansing and Analysis in Data Sciences. It explains how businesses often collect large volumes of data to make key decisions. Having such a priority for a company there is few to no space of making mistakes or bad decisions. An opposing force regarding this scenario are incorrect or invalid values filled in data cells. The authors describe many ways these values can be discovered, inspected and changed with validation. These may contain a human being on an end or a fully automated system.
Invalid values may occur due to reasons such as human error, invalid editing without validation, missing or extra values within the collection and wrong automated collection such as wrong SQL data extraction or masking. Looking at existing data cleaning techniques, the authors describe the following observations:

- Human Involvement
- Using an Automated tool
- Semi Automated approach

Every method has its own pros and cons. **Human Involvement** requires data cleansing at the hand of a professional person who understands as much there is about the domain specially the subject at hand. Though budget consuming and requiring the dispensing of man power this process provides slim chances to occur for such a scenario, nonetheless *a human being is a human after all* and can makes mistakes. **Automated tools** removes the need for human interaction completely and shows promising results, but presents a serious hold up in case of a deadlock such as in a SQL database with columns dispensing on each other for changes, in this it will requires a human presence, this is where **Semi Automated approach** comes in providing an assist by human with most of the progress being handled by a tool such as the data cleansing system *NADEEF and KATARA* as directed by another author cited in the paper.

The paper emphasizes on an architecture for data cleansing with data user presenting error to the analyst which is then repaired and validated by an expert user, while character human expertise as:

1. Detection - The detection of invalid cells
2. Repairing - Reported errors should be fixable by humans
3. Validation - Expert human analysis for validation on reparing
4. Specification - Humans are enabled to write specifications or a set of rules for detecting and repairing data errors

$$Expertise = \frac{correct}{validated} \tag{1}$$

Task Allocation is an important part of data cleansing where an automated tool assigns human experts with related tasks out of the machines domain or an available team of human experts do so among themselves, which then again presents a case of limited budget. **Identification of bottlenecks** represents specifying invalid values correctly with marking a score on each valid repair.

$$Quality = \frac{correct}{validated} \tag{2}$$