Assignment 2: Data Cleaning

Deadline: 16.06.2018
Presentation: 19.06.2018[1]

[1]The 800 Pound Gorilla in the Corner: Data Integration
Technische Universität Berlin

Abstract. *The goal of this assignment is to detect and correct data errors in a dataset, which is called data cleaning.*
Notice that:

- *The short questions usually require short answers too.*
- *One submission per each group is enough.*
- *You are free to use whatever tools/programming languages you intend to.*
- *Put your names and group number in the beginning of your final PDF report.*
- *Submit a compressed file that contains (1) the PDF report and (2) your source code files.*

1. Setup

1.1. Dataset

We have a relational address dataset with the following schema:

$$S(RecID, FirstName, MiddleName, LastName, Address,\\ City, State, ZIP, POBox, POCityStateZip, SSN, DOB).$$

The dataset has been attached to this document.

1.2. Evaluation

The cleaned version of this dataset, which is the ground truth, is not available for you! However, you can always send your cleaned dataset to our web service and get feedback to evaluate your data cleaning process. To communicate to the web service, you can either write your own script or use the attached *web_client.py*. This piece of code simply sends your dataset to the web service and gets the results. Below is a possible outcome:

```
$ python web_client.py
Number of dirty cells: 388126 (Number of erroneous cells in the data)
-----------------Error Detection Results-----------------
Number of detected cells: 1 (Number of changed values)
Number of Correctly Detected cells: 1 (cell was correctly identified as an
    error)
Detection precision: 1.0 (ratio of correctly detected cells over all detected
    cells)
Detection recall: 2.57648289473e-06 (ratio of correctly detected cells over all
     erroneous cells in the data)
Detection F1: 5.15295251297e-06
----------------Error Correction Results----------------
Destroyed clean cells: 0 (cell was correct but has been transformed into a
    wrong value)
```

```
Wrongly cleaned cells: 0 (cell was wrong but the cleaning was also not correct)
Undetected cells: 388125 (cell was erroneous but was not touched)
Number of cells that need yet to be cleaned: 388125 (sum of the 3 cell types
    above)
Correction precision = 1.0 (ratio of correctly corrected cells over all changed
     cells)
Correction recall = 2.57648289473e-06 (ratio of correctly corrected cells over
    all erroneous cells in the data)
Correction F1: 5.15295251297e-06
```

To use the *web_client.py* script, notice that:

- You might have to install *Python 2.7* and the *requests* module.
- The script assumes that the name of your dataset is *inputDB.csv* and it exists in the same directory.
- Opening and saving the CSV dataset file with spreadsheet tools (e.g. Microsoft Excel) might change encodings of the CSV file. It can leads to web server misbehavior or malfunction.
- The output distinguishes between error detection and error correction. Whenever you change a value in a cell, the evaluator in the server side will assume that you detected an error in that cell, which is error detection. On the other hand, if you transform the value into the actual correct value, you also committed an error correction.

2. Tasks

There are some desired data quality constraints for the dataset:

1. All alphabetical characters in all columns should be capitalized.
2. Address data should be compatible to the standard in `https://tools.usps.com/go/ZipLookupAction!input.action`.
3. The *State* column should contain the correct two character US state code.
4. *City* column should contain real city names.
5. *ZIP* column should be formatted as a 5 digit value.
6. *SSN* column should contain an 8-10 digit value.

Note that, you do not need to fix the *DOB*, *POBox*, and *POCityStateZip* columns.

2.1. Task 1: Error Detection

Here, we want to only *detect* data errors. To mark a cell as data error, you just need to change its value into something else.

1. Which of the mentioned data quality constraints can help you to detect data errors? How?
2. Report your best error detection precision, recall, and F1.

2.2. Task 2: Error Correction

Here, we want to not only detect data errors, but also automatically *correct* them to the true values. Therefore, it is not enough to just mark a cell as data error, you also need to update it to the correct value.

1. Which of the mentioned data quality constraints can help you to correct data errors? How?
2. Report your best error correction precision, recall, and F1.