Vishal Dubey

NIH-NIA Data Science Fellow

Civic Digital Fellowship Fall 2020

Under the guidance of Dana Plude, Partha Battacharya, and Jonathan King, we developed a synthetic dataset and comparison mechanism for different hashing and linking programs. Linkages among datasets is a key and necessary component of data science work as it allows one to identify subjects and data points across a multitude of areas. However, in some cases, there are errors in identifying these subjects due to inconsistent hashing as a result of random variability and inherent noise in the dataset itself. With many approaches involving hashing, the process whereby identifiable tokens are generated based on some input variables, there is not a general consensus on which hashing approaches provide the lowest error rate and highest rate of consistent linkages across a noisy dataset.

In order to compare the linkages and effectiveness across a multitude of platforms, it was first necessary to create a synthetic and realistic dataset, mirroring some of the noise that would be found in actual datasets. To do this, we began by first generating a complete synthetic dataset based on combining CRISP Legacy Data of NIH grant applications from the online available datasets for the past 39 years on the NIH Exporter website. Combining this data, resulted in a total of approximately 2.6 million datapoints. We first dropped any rows which had empty organization names, addresses, or first, middle, and last names. After doing these drops, we had roughly 300,000 unique persons in our dataset, equating to around 2.1 million rows when considering duplicates.

In order to simulate the actual datasets used to generate tokens, we found that it was necessary to include location information such as zip code, city, and state as well. Through the Python library “uszipcode,” we were able to cross reference the state codes in the downloaded data with zip codes and cities in the appropriate state to impute the values. In this process, we ensured that duplicates of the same person within an organization would retain the same location values as well. For the approximately 8% of the dataset which lacked associated state codes, we imputed the values by selecting random location values within the state associated with the given institution.

Upon adding location history variables, we then also had to add columns for day, month, and year of birth, being DOB, MOB, and YOB respectively. Assuming a normally centered distribution of age with the median government age as 39 years old, we generated birthdays for each unique person. In order to simulate reality as close as possible, we chose to add gender as a column through using a natural language processing library in Python named “gender-guesser detector model.” Finally, we imputed SSN by creating random ones and caching each result to ensure that each person has a unique identifier. This would serve as our GIID. This completed the sample dataset, resulting in approximately 300,000 unique persons within a collection of 2.1 million datapoints.

The next step was to impute noise and variation into this dataset and to accomplish this, we deployed a five-pronged approach. First, we selected the 5 areas in which to manufacture variation, ultimately selecting zip code, gender, year of birth, first name, last name, and city. For zip code, we randomly changed the current code by a 0 to 200 value and for city, we randomly sampled another city within the same state using the Python library mentioned earlier, both on 10% of the dataset. For gender, we randomly reassigned gender with an equal chance of reassignment for any given person in our 8% sample. First and last name were trickier as we had to consider key stroke errors in typing names which would happen. The different cases in this regard were either a letter would be typed around the given one or that there would be an insertion or deletion. For both of these cases, we created our own library of possible key stroke errors for each given letter and then randomly imputed the noise in a 4.5% sample. Throughout this process, we ensured that the GIID for each unique person remained the same, so as to reidentify them upon receiving linkage results from Datavant, Acumen, and NIH CIT.