Service Records integration with master addresses

What is the problem you want to solve?

For one operation unit, all service installation records are stored and maintained in a legacy file store with address info. However, no system or field linkage is built between this file store and ERP application. The addresses are entered manually each time the file added to the file store. The only way to find the records for an address in ERP is to manually search the file store using the address.

Who is your client and why do they care about this problem? In other words, what will your client do or decide based on your analysis that they wouldn’t have done otherwise?

My client is a natural gas company. Service pipeline operations (install, replace, repair, …) are core business functions to provide customers service in a safe, efficient manner. To integrate the legacy records into assets management system, matching the addresses in legacy file store between the master addresses of ERP will be the first step and the only way to bridge the gap. Manual data mining and matching will be time consuming and almost mission impossible for over million records. A systematical solution to resolve this problem is required.

What data are you using? How will you acquire the data?

The data I will use for this project comes from both ERP master address data set and legacy file store address data set. The data and result will be exported into csv files and will then be imported into the Python notebook.

Briefly outline how you’ll solve this problem. Your approach may change later, but this is a good first step to get you thinking about a method and solution.¶

My approach to solving this problem is to get all the required data into a centralized staging database and then perform initial analysis, merge/clean the data. and use a third-party address tool for initial matching. I will then import the data into notebook, perform Data Wrangling / EDA, and test different NLP algorithms to determine which works best for the problem I am trying to solve.

What are your deliverables? Typically, this includes code, a paper, or a slide deck.

My deliverables will be a jupyter notebook and slide deck that will be published to my github account. I will include all my findings and all documented python code.