CAS 764: Final Project Progress

For the final project, I started to get research based tools. I downloaded 4 of them: Magellan, SERF, FRIL and Febrl. However, most of them require additional libraries.

1. Magellan: It's python based system and requires 10-12 external libraries. It was easy to get external libraries because most of the are up to date. After downloading it, I wanted test it. I looked at its website and found step-by-step instructions. It doesn't have any UI so it only works with python commands. First, I tried to use it with python commands then, I created a python script for the same job. Magellan has its own data for examples and it works perfectly on them. However, I encountered some errors while using it with other datasets. Later, I fixed some of errors for 1 dataset but another dataset got different errors. I am going to check and fix them later. Also, I tried sampling, blocking (1+ blockers and debugging) and matching (using 1 matcher, selecting between 6 of them, debugging) operations.
2. FRIL was the easiest system. I downloaded it form a website and that was it. It only requires a Java and a specific system variable. After the requirements, I tried to use it. It has a basic UI but the others don't even have that, so it's easier to use than all of the others. FRIL has 2 different modes: Linkage and Deduplication. Linkage takes 2 tables and Dedup takes 1 table. Both of them consists of 2-3 steps and user can select various techniques and parameters to get the result. At the endof the test, I got a sample file which contains the duplicated tuples.
3. SERF is similar to FRIL. It only requires a Java and it works with Java commands. I tried to test it however, its run command requires 2 different files.
   - *java -cp "../libs/secondstring-20030401.jar:../serf.jar" serf.ER example.conf*

   One of them is in the downloaded zip file but the other one (serf.ER) isn't in the file. When I tried to run it with that command, I got missing file errors for (serf.ER) file. I am going to check this error later.
4. Febrl was the hardest one. It is a python based system and it requires a few external libraries/programs. I think Febrl is an old system so its requirements are hard to get. For example, it requires PyGTK library. I tried to get PyGTK but I couldn't download it from pip. I tried different methods but couldn't solve the problems yet. If I cannot solve the problems, I might find a different system.

Since commercial tools have limited time usage, I didn't get them yet. Firstly, I would like to use the free research based tools and gain some experience about how to use entity matching systems. After that, I am going to get free trial versions of some of the entity matching systems like DataLadder and SAS DataFlux and I am going to test them with the same data. However, I think some of the commercial systems have number of data limitation (eg. 1000). In that case, I may need to shrink the datasets.

2019-02-24

Baran Kaya, kayab@mcmaster.ca, 400284996