Homework 0: Exploring Entity Resolution

Due: Friday February 5, 2021 at 5 PM EDT

General instructions for homeworks: Your code must be completely reproducible and must compile. No late homeworks will be accepted.

Reading Read the paper Binette and Steorts (2020) to get an overview of entity resolution. You'll want to refer to this during the course of the semester as it's meant to be a quick reference regarding the concepts that we will be covering. For more details, refer to the book by Christen (2012).

Advice: Start early on the homeworks and it is advised that you not wait until the day of as these homeworks are meant to be longer and treated as case studies.

Commenting code Code should be commented. See the Google style guide for questions regarding commenting or how to write code https://google.github.io/styleguide/Rguide.xml.

R Markdown Test

0. Open a new R Markdown file; set the output to HTML mode and "Knit". This should produce a web page with the knitting procedure executing your code blocks. You can edit this new file to produce your homework submission.

Total points on assignment: 5 (reproducibility) + 25 points for the assignment.

1. (15 points) Find **three examples** of entity resolution tasks using information publicly available. For each of the three examples, describe the problem at hand and what information is needed in order to distinguish the unique entities (or link/merge record that belong to each entity). Be sure to include a list of the records for each example and the public source where you found them. **Entities can be people or objects. One example can be about yourself if you find this easier to discuss and talk about.**

   For example, you might use Google Scholar to search for a textbook. What kind of results do you get back? Are they unique or not? As another example, what happens when you try and search for someone on PubMed (such as an author or paper). How could identify unique papers? What other public records might have linkage problems?

2. (5 points) Download and install the `RecordLinkage` package in R on your computer. View the help documentation for `strcmp`. Which string metrics can this function calculate?

3. (5 points) Read the six page paper describing the Jaro-Winkler string metric, "String Comparator Metrics and Enhanced Decision Rules in the Fellegi-Sunter Model of Record Linkage. Concentrate on Sections 1, 2.3, 2.4, and 3. We will discuss the Fellegi-Sunter model in more depth in class.

   On page 4, Table 1 provides Jaro-Winkler similarity values for several pairs of words. Pick a pair of words and use the function `jarowinkler` to experiment with typographical errors in different places.

   What happens when you switch letters at the beginning of the word? the end? What if the letter moves two or three places? What happens if you leave out letters in one of the words? What types of changes

result in larger decreases in the similarity score? Describe your results and provide reproducible code in order to back up any claims.