36-491 HOMEWORK 1

Due: Wednesday 9/4/13

1. Find three examples of record linkage problems using information available publicly. Describe what the problem is and what information you need to be able to distinguish the unique entities or link up the records belonging to each entity. Include a list of the records in question (copy the list, grab a screenshot, etc).

   Entities can be people or objects, but one example should be about yourself.

   For example, you might use GoogleScholar to search for a specific textbook - what kind of results do you get back? How do they match up? Do they? Or finding records with your name? Do they all refer to you? How could you identify yours? What other mixups occur in public records? What happens when searching for someone on Facebook? How do you distinguish different people? Be creative. Don't just use Google. What other public records might have linkage problems?

2. Download and install the RecordLinkage package in R on your computer. Note that you will need to download and install several other packages in this process (R will tell you which ones).

   Also, the necessary package *e1071* is temperamental; it may or may not load depending on its version and your R version. For example, on my laptop, it loads on my 32-bit R but not my 64-bit R. If you have trouble, contact Sam.

   View the help documentation for **strcmp**.
   Which string metrics can this function calculate?

3. 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 detail later.

   On p.4, Table 1 has 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.