Evaluation Metrics for Blocking/Entity Resolution

STA 325: Homework 2

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 (2022) 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: 2 (reproducibility) + 23 points for the assignment = 25 total points.

1. (4 points) What are the four main challenges of entity resolution?

2. (4 points, 1 point each) Suppose there are 10 records in a data set. a.) What are the total number of brute-forte comparison needed to make all-to-all record comparisons? b.) Repeat this for 100 records, 1000 records, 10,000 records. c.) What do you observe about the number of comparisons that need to be made?

3. (9 points) Consider the following record linkage data set with 1,000,000 total records that are matched between two databases. Assume that 500,000 are true matches. Assume a classified (or method) finds 600,000 record pairs as matches, and of these 400,000 correspond as true matches. The number of TP + FP + TN + FN = 50,000,000.

a. (4 points) Given the information above, find the following information in the confusion matrix: TP, FP, TN, and FN.
b. (1 point) Calculate the accuracy. Comment on the reliability of this metric for this problem.
c. (1 point) Calculate the precision.
d. (1 point) Calculate the recall.
e. (1 point) Calculate the f-measure.
f. (1 point) Comment on the reliability of the precision, recall, and f-measure for this problem.

4. (6 points) We will revisit the Italian Survey on Household and Wealth (SHIW) from class, which is a sample survey 383 households conducted by the Bank of Italy every two years (2008 and 2010). The data set is anonymized to remove first and last name (and other sensitive information).

a. (0 points) Please load the data set in the way that we did in class and block based upon gender.
b. (1 point) Plot the size of the blocks and comment on how many their are and their relative size.
c. (1 point) Calculate the reduction ratio and interpret its meaning.

d. (2 points) Calculate the precision and recall. Interpret the meaning of each.
e. (1 point) Would this be a reasonable approach for blocking. Explain.
f. (1 point) Would blocking on gender be recommended for entity resolution. Explain.

```
library(italy)
library(assert)
data(italy08)
data(italy10)
knitr::opts_chunk$set(echo = TRUE,
                      fig.width=4,
                      fig.height=3,
                      fig.align="center")
head(italy08)
```

```
##        id PARENT SEX ANASC NASCREG CIT ACOM4C STUDIO Q QUAL SETT IREG
## 1 1040021      1   2  1948      16   1      0      5 1    2    3   16
## 2 1040022     10   2  1952      16   1      0      7 1    2    3   16
## 3 1110521      1   1  1972      20   1      2      5 1    1    4   20
## 4 1110522      3   1  1935      20   1      2      2 3    6    5   20
## 5 1110523      3   2  1941      20   1      2      3 3    6    5   20
## 6  119401      1   1  1941       7   1      0      4 3    6    5    7
```

```
head(italy10)
```

```
##        id PARENT SEX ANASC NASCREG CIT ACOM4C STUDIO Q QUAL SETT IREG
## 1 1040021      1   2  1948      16   1      0      5 3    6    5   16
## 2 1040022     11   2  1952      16   1      0      7 1    2    3   16
## 3 1110521      1   2  1941      20   1      2      3 3    6    5   20
## 4 1110522      2   1  1935      20   1      2      2 3    6    5   20
## 5 1110523      6   1  1972      20   1      2      5 1    1    4   20
## 6  119721      1   2  1948      16   1      2      2 2    5    4   17
```

```
id08 <- italy08$id
id10 <- italy10$id
id <- c(italy08$id, italy10$id) # combine the id
italy08 <- italy08[-c(1)] # remove the id
italy10 <- italy10[-c(1)] # remove the id
italy <- rbind(italy08, italy10)
head(italy)
```

```
##   PARENT SEX ANASC NASCREG CIT ACOM4C STUDIO Q QUAL SETT IREG
## 1      1   2  1948      16   1      0      5 1    2    3   16
## 2     10   2  1952      16   1      0      7 1    2    3   16
## 3      1   1  1972      20   1      2      5 1    1    4   20
## 4      3   1  1935      20   1      2      2 3    6    5   20
## 5      3   2  1941      20   1      2      3 3    6    5   20
## 6      1   1  1941       7   1      0      4 3    6    5    7
```