Homework 2

Nathan Yang

9-9-2024

Exercise 1

The 4 main challenges of Entity Resolution are

1. Costly manual labelling
2. Scalability/computational efficiency
3. Limited treatment of uncertainty
4. Unreliable evaluation

Exercise 2

a. For 10 records, there are 10^2 or 100 total brute-force comparisons
b. For 100 records, there are 100^2 or 10,000 total brute-force comparisons
   For 1000 records, there are 1000^2 or 1,000,000 total brute-force comparisons
   For 10000 records, there are 10000^2 or 100,000,000 total brute-force comparisons

c. The number of comparisons grows quadratically with the number of records.

Exercise 3

Dataset with 1,000,000 entries, 500,000 are true matches, method found 600,000 as matches, and 400,000 of these are true matches. TP + FP + TN + FN = 50,000,000

a. TP = 400,000, FP = 200,000, TN = 49,300,000, FN = 100,000
b. Accuracy $= \frac{400{,}000+49{,}300{,}000}{50{,}000{,}000} = 0.994$
c. Precision $= \frac{400{,}000}{400{,}000+200{,}000} = 2/3$
d. Recall $= \frac{400{,}000}{400{,}000+100{,}000} = 0.8$
e. F-Measure $= \frac{2*(2/3)*0.8}{(2/3)+0.8} = 0.7\bar{2}$
f. Precision, recall, and f-measure are much better metrics than accuracy because there is a large number of true negatives which leads to a class imbalance.

Exercise 4

Italian Household Survey on Household and Wealth

a.
```r
# Load necessary packages
if (!require("pacman")) {
install.packages("pacman")
library(pacman)
}
```

```
## Loading required package: pacman
```

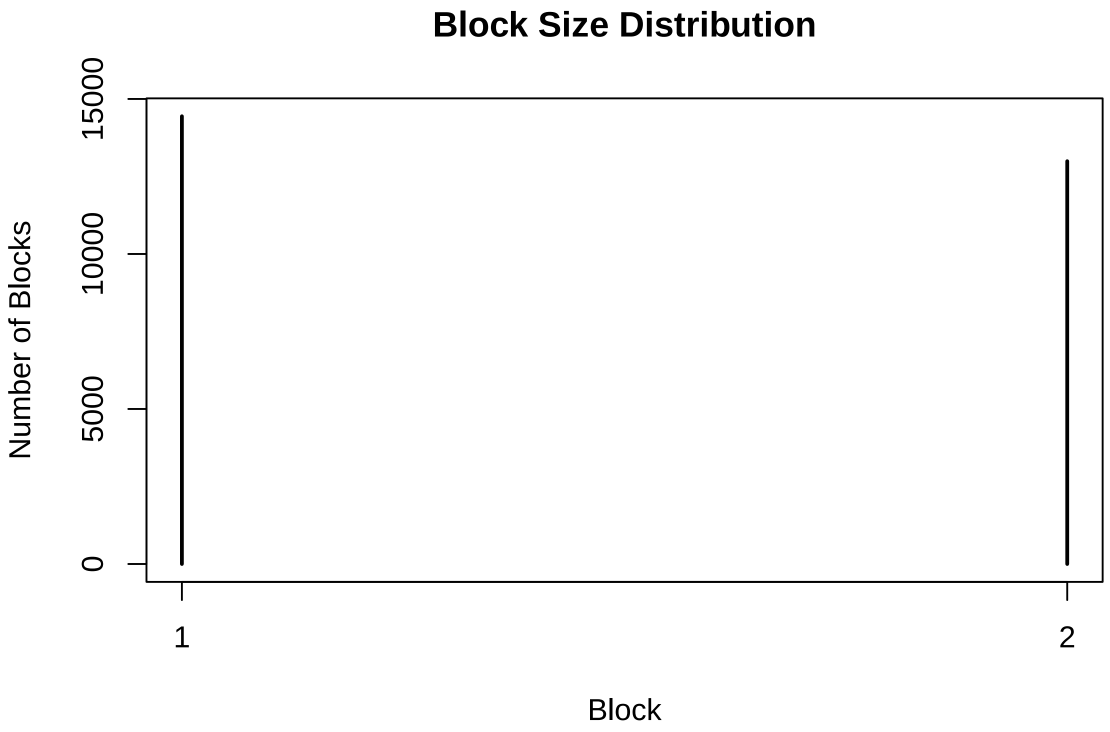
```
p_load(RecordLinkage, blink, italy, tidyverse, assert)

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
```

```
blockByGender <- italy$SEX
recordsPerBlock <- table(blockByGender)
```

b.

```
# Plot the blocks
plot(recordsPerBlock, xlab = "Block", ylab = "Number of Blocks", main = "Block Size Distribution")
```



There are only 2 blocks and Block 1 is slightly larger than Block 2 with 14442 and 12993 records respectively.

c.

```r
# Function to calculate reduction ratio
ReductionRatio <- function(dataset) {
  n_all_comp = choose(length(dataset), 2)
  n_block_comp = sum(choose(table(dataset), 2))
  (n_all_comp - n_block_comp) / n_all_comp
}

ReductionRatio(blockByGender)
```

```
## [1] 0.4986234
```

The reduction ratio is 0.50. We reduced the comparison space by roughly 50%.

d. Precision: 3.6e-05
.0036% of the classified matches were true matches
Recall: 0.91
91% of the true matches are classified correctly

```r
# Precision Function
precision <- function(block.labels, IDs) {
  ct = xtabs(~block.labels+IDs)
  # Number of true positives
  TP = sum(choose(ct, 2))
  # Number of positives = TP + FP
  P = sum(choose(rowSums(ct), 2))
  return(TP/P)
}

# Recall Function
recall <- function(block.labels, IDs) {
  ct = xtabs(~IDs+block.labels)
  # Number of true positives
  TP = sum(choose(ct, 2))
  # Number of true links = TP + FN
  TL = sum(choose(rowSums(ct), 2))
  return(TP/TL)
}

precision(blockByGender, id)
```

```
## [1] 3.599727e-05
```

```r
recall(blockByGender, id)
```

```
## [1] 0.9113109
```

e. This is an okay approach to blocking because 91% of the true matches are classified correctly (high recall)

f. This is not a recommended approach to entity resolution because a very very minute percentage of the classified matches are true matches (very low precision)