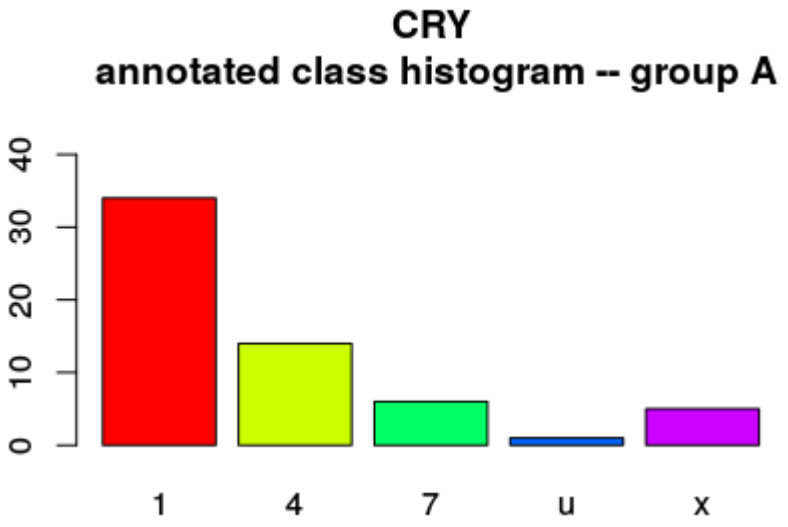
A gentle tutorial on manual annotation data analysis

NPFL 054 lab session (Hladká and Holub, 2016)

Part I -- Reading data into data frame

```
### reading the data
> cry.A = read.csv("cry-A.csv", header = F, sep = ";",
                   col.names = c("id", "class", "void"))

### structure of the data frame
> str(cry.A)
'data.frame':60 obs. of  3 variables:
 $ id   : int  28873523 8410635 25049966 4987269 26009795 27200971 15914329 ...
 $ class: Factor w/ 5 levels "1","4","7","u",..: 4 1 1 5 5 2 3 2 1 1 ...
 $ void : logi  NA NA NA NA NA NA ...
>

### remove the void column
> cry.A$void = NULL

> str(cry.A)
'data.frame':60 obs. of  2 variables:
 $ id   : int  28873523 8410635 25049966 4987269 26009795 27200971 15914329 ...
 $ class: Factor w/ 5 levels "1","4","7","u",..: 4 1 1 5 5 2 3 2 1 1 ...
>

### check the number of unique sentences
> length(unique(cry.A$id))
[1] 60
>

### the 'class' distribution
> table(cry.A$class)

 1  4  7  u  x
34 14  6  1  5
>

### histogram
> barplot(table(cry.A$class), main="CRY\n annotated class histogram -- group A",
          col = rainbow(5),
          ylim = c(0, max(table(cry.A$class))+10))
```



Part II -- Joining data using data tables

```
*** Our recommendation: get familiar with the great R package data.table!
       https://cran.r-project.org/web/packages/data.table/data.table.pdf
       https://cran.r-project.org/web/packages/data.table/vignettes/datatable-intro.pdf

# load data.table package
> library(data.table)

### if the package has not been installed yet, install it first
### > install.packages("data.table")


### transform data frame to data table
> cry.A = data.table(cry.A)

> tables()
     NAME  NROW NCOL MB COLS     KEY
[1,] cry.A   60    2  1 id,class
Total: 1MB
>

### data table can be directly read from a file
> cry.B = fread("cry-B.csv", header = F, sep = ";",
                 select = c(1,2), col.names = c("id", "class"))

### set index to the id columns
> setkey(cry.A, "id")
> setkey(cry.B, "id")

> tables()
     NAME  NROW NCOL MB COLS     KEY
[1,] cry.A   60    2  1 id,class id
[2,] cry.B   60    2  1 id,class id
Total: 2MB
>

### join the tables using the common index
> cry.AB = cry.A[cry.B]
> setnames(cry.AB, c("id", "A", "B"))

### now simply make a table with the IAA
> table(cry.AB[, c(2,3), with=F])
   B
A    1  4  7  u  x
  1 25  3  1  4  1
  4  4  6  2  2  0
  7  1  1  3  1  0
  u  0  1  0  0  0
  x  1  0  2  1  1
>
```

Part III -- Exercises

***** Exercise A)**
Read annotation data and compute the Cohen's kappa value between groups A and B.

Help: Use table(cry.A$class) to get frequencies of the labels used by group A.

Correct answer
```
Pr(a) = 0.5833333
Pr(e) = 0.3538889
kappa = 0.3551161
```

***** Exercise B)**
Read the gold-standard data (= 250 examples) and the output of the automatic classifier F1 (= the same 250 instances with labels assigned by F1 predictor).
What is the classifier accuracy?
Display the confusion matrix.
Then transform the confusion matrix into percentages.

Help: Confusion matrix is like the IAA matrix for F1 and GS.
Using table(., .) you can easily get it.
To do an operation with every column of a matrix, learn the apply() function.

Confusion matrix

```
   GS
F1    1   4   7   u   x
  1 125   9   0   6   8
  4   2  48   0   2   1
  7   0   0  12   1   0
  u   4   1   1  22   0
  x   0   1   0   2   5
```

Sample accuracy is 0.848.

Probability of errors in columns – rounded percentages

```
   GS
      1    4    7    u    x
  1 95.4 15.3  0.0 18.2 57.1
  4  1.5 81.4  0.0  6.1  7.1
  7  0.0  0.0 92.3  3.0  0.0
  u  3.1  1.7  7.7 66.7  0.0
  x  0.0  1.7  0.0  6.1 35.7
```