spamizer

Marc Sànchez, Francesc Xavier Bullich, Gil Gassó

5/8/2019

Estudi de les variables PHI i K

El que es pretén és realitzar un estudi de quan les variables phi i k considerades com a constants en l'execució del programa es comporten de manera adient per el filtratge.

TODO : Explicar filtres. Stanford.

TODO : Explicar kfold.

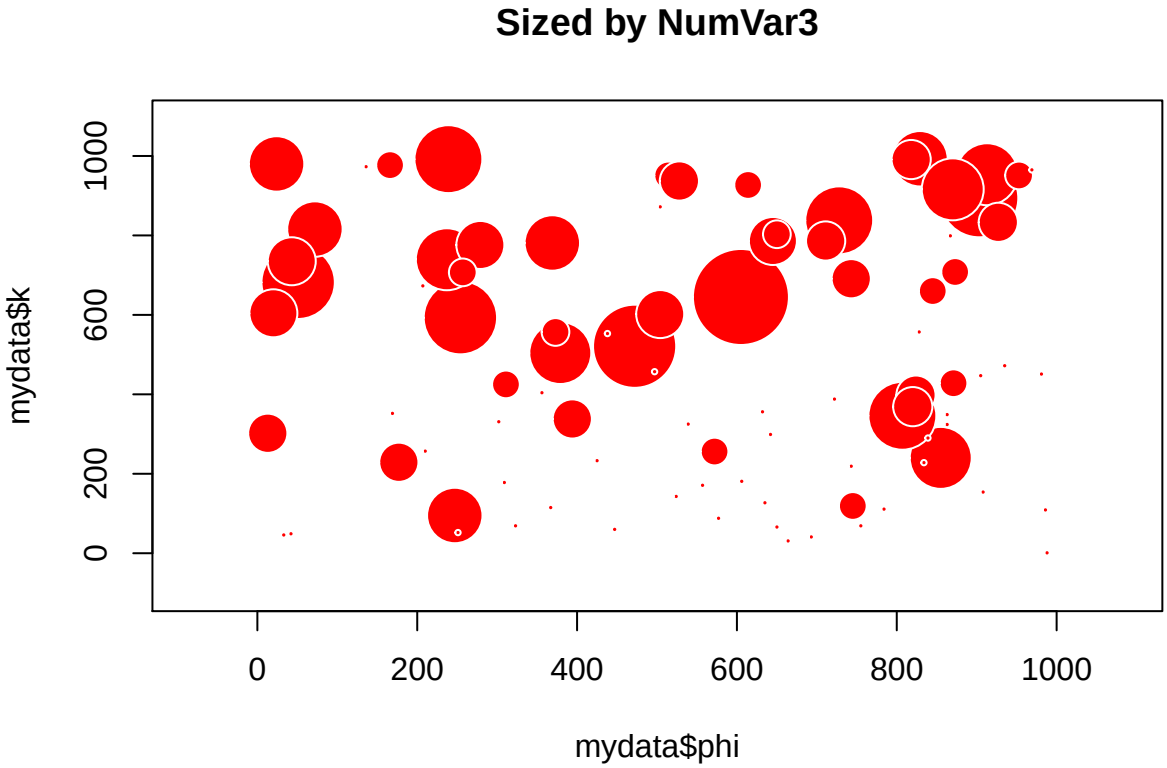
Execució amb 200 mails i 100 iteracions

En el següent gràfic la grandària dels punts estipula quant de gran és l'error no desitjat, és a dir, quan un correu considerat **HAM es filtra com SPAM**. Als eixos hi podem veure els valors de phi i k utiltizats per a la validació. El percentatge de correus utilitzats sobre els 200 correus totals és d'entre 5% i 15% i la selecció d'aquest valor és aleatòria.

```
# Carreguem les dades per a l'execució del gràfic.
mydata = read.csv("/Users/marcsanchez/Projects/spamizer/analisys/200m-100n.csv")
head(mydata)
```

```
##   id phi   k TP TN FP FN
## 1 48 472 521  2  4  9  3
## 2 49 855 240  4  5  5  1
## 3 50  51 682  1  3  7  2
## 4 51 745 119 13 12  1  3
## 5 52 807 346  6 13  6  3
## 6 53 693  41 11  1  0  5
```

```
radius <- sqrt(mydata$FP/pi)
symbols(mydata$phi, mydata$k, circles = radius, inches = 0.25, fg = "white",
    bg = "red", main = "Sized by NumVar3")
```



Referències

- R graphics