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.

Anàlisi de la PHI i la K

PHI

Si tenim en compte el què representa ela valors de phi, el que ens trobem és que la phi és el factor d'increment de la probabiltat per que un correu sigui considerat SPAM. És a dir un valor de phi = 2, provoca que per que un correu sigui considerat spam ha de ser 2 cops superior a la probabiltiat de que sigui ham. Un valor de phi = 1 fa que no hi hagi increment obligatori per a la comparació.

El valor mínim que té sentit assignar-li a phi és 1 i el màxim el podríem limitar a 5 com a molt o inclús a 6 si el que volem és no tenir cap correu que sigui Ham i que el consideri com Spam.

K

Quan apliquem el suavitzat hem de tenir en compte que donats el bag of words de ham i el de spam, què passa si la paraula no existeix? doncs que el valor de les multiplicacions serà 0 i farà que si una paraula no existeix aquesta paraula ens determini si un correu és ham o és spam.

Per tant la k estipula el valor que se li assigna a una paraula quan aquesta no és present. Aquest valor no pot ser 0 peró pot ser proper a zero. Si fos zero es provocaria el mateix cas que l'esmentat anteriorment. Tanmateix no té sentit aplicar un valor molt gran a la k ja que si ho fem aquest valor provocaria que les paraules que no existeixen fossin puntuades molt altes i que les aparicions no computessin tant.

Limitarem els valors de k en un rang de (0 - 3].

Exemple de funció K

```
#Calculem els tcr dels valors
# BASE :  (NSPAM) / (50 * NHAM + NSPAM)
base <- results$NSPAM / (50 * results$NHAM + results$NSPAM)
# WERR: (50 * FP + FN)/(50 * NHAM + NSPAM) + 0.000001
werr <- (50 * results$FP + results$FN) / (50 * results$NHAM + results$NSPAM) + 0.000001
# TCR : BASE / WERR
tcr <- base/werr
```

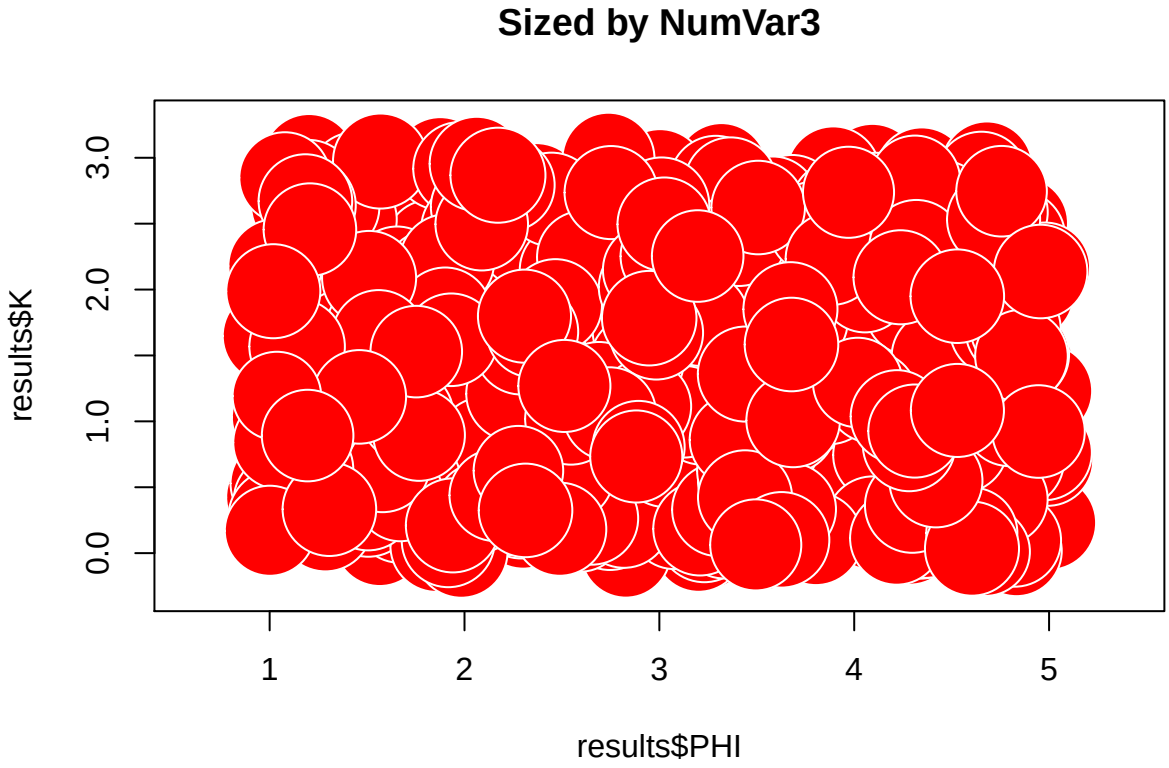
```
library(scatterplot3d)
scatterplot3d(x=results$PHI, y=results$K, z=tcr)


# Cargar librerias
library(ggplot2)
# Generar la matriz
valores <- data.frame(results$PHI, results$K, tcr)
names(valores) <- c("phi", "k", "tcr")
head(valores)
# Pintar
p <- ggplot(valores, aes(x = phi, y = k))
p + geom_tile(aes(fill = tcr))
```

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.

```
head(results)
```

```
##    ID      PHI        K   TP   TN FP FN NHAM NSPAM
## 1 305 4.040728 2.451330 1047 1024  7 79 1054  1103
## 2 306 3.680707 2.326060  847  819  5 44  852   863
## 3 307 4.088587 1.293652 1456 1508  8 71 1464  1579
## 4 308 1.842563 2.567766  925  825  4 47  929   872
## 5 309 1.195359 1.544295 1095 1050  6 43 1101  1093
## 6 310 2.171663 1.529208 1293 1220 10 65 1303  1285
```

```
radius <- sqrt((1-(results$FP/100))/pi)
symbols(results$PHI, results$K, circles = radius, inches = 0.25, fg = "white",
    bg = "red", main = "Sized by NumVar3")
```



Referències

- R graphics