Almost Famous: Verify Spam Classification Prediction

Cindy Lamm

01:48, Thursday 22nd January, 2015

Load prediction results:

```
prediction <- read.table("out/unsupervised/train_kmeans_prediction.csv",
                         col.names=c("visit_id", "uid", "is_spam"),
                         colClasses="factor")
str(prediction)

## 'data.frame': 204404 obs. of  3 variables:
##  $ visit_id: Factor w/ 173430 levels "10001576436",..: 17629 17629 15982 15982 15982 71655 71655 7165
##  $ uid     : Factor w/ 169391 levels "100001489","10000221",..: 165764 165764 37251 37251 37251 1981
##  $ is_spam : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 ...

table(prediction$is_spam)

##
##      0      1
## 201671   2733

prop.table(table(prediction$is_spam))

##
##          0          1
## 0.98662942 0.01337058
```

Load spam ids:

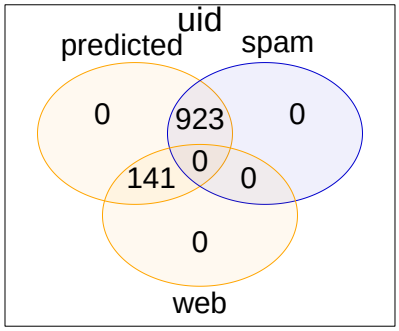
```
spamIdFile <- "out/unsupervised/spam_numeric_id.csv"
spamIds <- read.table(spamIdFile, header=TRUE, sep=" ", colClasses="factor")
str(spamIds)

## 'data.frame': 4404 obs. of  2 variables:
##  $ visit_id: Factor w/ 1482 levels "10199862810",..: 146 146 130 130 130 602 602 602 602 1409 ...
##  $ uid     : Factor w/ 1060 levels "100191","100547",..: 1038 1038 238 238 238 9 9 9 9 320 ...
```

Load web ids:

```
webIdFile <- "out/unsupervised/web_numeric_id.csv"
topX <- nrow(prediction)-nrow(spamIds)
webIds <- read.table(webIdFile, header=TRUE, sep=" ", colClasses="factor", nrows=topX)
str(webIds)
```

```
## 'data.frame': 200000 obs. of  2 variables:
##  $ visit_id: Factor w/ 171948 levels "10001576436",..: 14760 4138 171295 21028 40751 45720 3479 7006
##  $ uid     : Factor w/ 168331 levels "100001489","10000221",..: 138508 147095 2074 115956 75610 11897
```

Plot Venn diagram:



Something is off - the counts don't add up!!!

```
print(vc)
```

```
##   predicted spam web Counts
## 1        NA   NA  NA     NA
## 2         0    0   1      0
## 3         0    1   0      0
## 4         0    1   1      0
## 5         1    0   0      0
## 6         1    0   1    141
## 7         1    1   0    923
## 8         1    1   1      0
## attr(,"class")
## [1] "VennCounts"
```