Evaluation of Smelly London - implied smells analysis

Barbara McGillivray

28/02/2018

# Initialization

dir = file.path("/Users", "bmcgillivray", "Documents", "OneDrive", "OneDrive - The Alan Turing Institute", "Research", "Smelly\_London", "evaluation", fsep = "/")

## Read data

# Replace separator ":::" with "\t" in ModelScores.txt and save as ModelScores1.txt  
ev = read.csv(paste(dir, "ModelScores1.txt", sep = "/"), sep = "\t")  
dim(ev)

## [1] 1954 1442

Replace NAs in manual annotation column with zeros:

ev$X1[is.na(ev$X1)] <- 0

# Evaluation metrics

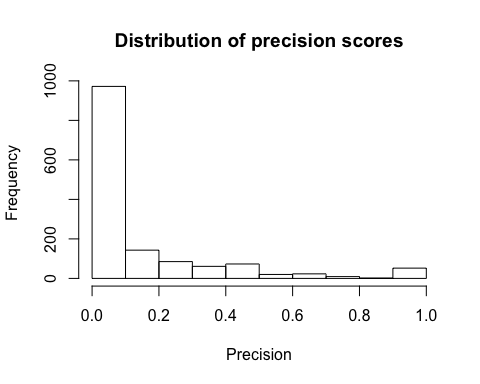
precision <- function(i) {  
 if (sum(ev[,i]) > 0){  
 p = (drop(ev[,2] %\*% ev[,i]))/sum(ev[,i])  
 }  
 else {  
 p = 0  
 }  
 p = round(p, 2)  
}  
recall <- function(i) {  
 r = (drop(ev[,2] %\*% ev[,i]))/sum(ev[,2])  
 r = round(r, 2)  
}  
precision\_scores = lapply(3:ncol(ev), precision)  
#names(precision\_scores) = colnames(ev)[3:ncol(ev)]  
recall\_scores = lapply(3:ncol(ev), recall)  
#names(recall\_scores) = colnames(ev)[3:ncol(ev)]  
eval = data.frame(model = colnames(ev)[3:ncol(ev)], precision = as.numeric(precision\_scores), recall = as.numeric(recall\_scores))  
eval$Fscore = round(ifelse(eval$precision+eval$recall>0, 2\*(eval$precision\*eval$recall)/(eval$precision+eval$recall), 0), 2)

# Analysis

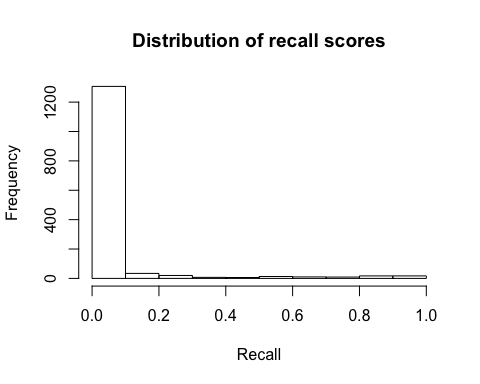
summary(eval)

## model precision recall   
## GenCBOW2x200x100dep : 1 Min. :0.0000 Min. :0.00000   
## GenCBOW2x200x100one : 1 1st Qu.:0.0000 1st Qu.:0.00000   
## GenCBOW2x200x100r.one: 1 Median :0.0000 Median :0.00000   
## GenCBOW2x200x100r.two: 1 Mean :0.1355 Mean :0.05378   
## GenCBOW2x200x100two : 1 3rd Qu.:0.1700 3rd Qu.:0.02000   
## GenCBOW2x200x200dep : 1 Max. :1.0000 Max. :0.97000   
## (Other) :1434   
## Fscore   
## Min. :0.00000   
## 1st Qu.:0.00000   
## Median :0.00000   
## Mean :0.02908   
## 3rd Qu.:0.03250   
## Max. :0.23000   
##

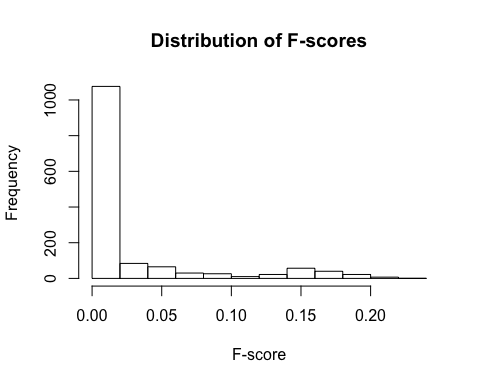
hist(eval$precision, main = "Distribution of precision scores", xlab = "Precision")



hist(eval$recall, main = "Distribution of recall scores", xlab = "Recall")



hist(eval$Fscore, main = "Distribution of F-scores", xlab = "F-score")



Maximum F-score:

eval[which.max(eval$Fscore),c("model")]

## [1] GenIndyDISTBOWx200x200two  
## 1440 Levels: GenCBOW2x200x100dep ... SpecSGBOWx20x50two

top\_p = eval[eval$precision == max(eval$precision),]  
top\_p = top\_p[order(top\_p$precision, decreasing = T),]  
top\_p = top\_p[order(top\_p$model, decreasing = F),]  
top\_r = eval[eval$recall == max(eval$recall),]  
top\_r = top\_r[order(top\_r$precision, decreasing = T),]  
top\_r = top\_r[order(top\_r$model, decreasing = F),]  
top\_f = eval[eval$Fscore == max(eval$Fscore),]  
top\_f = top\_f[order(top\_f$precision, decreasing = T),]  
top\_f = top\_f[order(top\_f$model, decreasing = F),]  
kable(top\_p)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | model | precision | recall | Fscore |
| 382 | GenCBOW2x200x100two | 1 | 0.01 | 0.02 |
| 261 | GenCBOW2x200x20one | 1 | 0.02 | 0.04 |
| 322 | GenCBOW2x200x50two | 1 | 0.01 | 0.02 |
| 23 | GenCBOW2x20x20r.one | 1 | 0.01 | 0.02 |
| 20 | GenIndyDIST2x20x20dep | 1 | 0.01 | 0.02 |
| 740 | GenIndyDIST5x200x20dep | 1 | 0.02 | 0.04 |
| 497 | GenIndyDIST5x20x20two | 1 | 0.01 | 0.02 |
| 560 | GenIndyDIST5x20x50dep | 1 | 0.01 | 0.02 |
| 1220 | GenIndyDISTBOWx200x20dep | 1 | 0.01 | 0.02 |
| 1100 | GenIndyDISTBOWx20x100dep | 1 | 0.01 | 0.02 |
| 977 | GenIndyDISTBOWx20x20two | 1 | 0.01 | 0.02 |
| 1040 | GenIndyDISTBOWx20x50dep | 1 | 0.01 | 0.02 |
| 312 | GenIndyNORM2x200x50two | 1 | 0.01 | 0.02 |
| 135 | GenIndyNORM2x20x100dep | 1 | 0.01 | 0.02 |
| 195 | GenIndyNORM2x20x200dep | 1 | 0.01 | 0.02 |
| 75 | GenIndyNORM2x20x50dep | 1 | 0.01 | 0.02 |
| 72 | GenIndyNORM2x20x50two | 1 | 0.01 | 0.02 |
| 855 | GenIndyNORM5x200x100dep | 1 | 0.01 | 0.02 |
| 792 | GenIndyNORM5x200x50two | 1 | 0.01 | 0.02 |
| 612 | GenIndyNORM5x20x100two | 1 | 0.01 | 0.02 |
| 675 | GenIndyNORM5x20x200dep | 1 | 0.01 | 0.02 |
| 552 | GenIndyNORM5x20x50two | 1 | 0.01 | 0.02 |
| 1332 | GenIndyNORMBOWx200x100two | 1 | 0.01 | 0.02 |
| 1395 | GenIndyNORMBOWx200x200dep | 1 | 0.01 | 0.02 |
| 971 | GenIndyNORMBOWx20x20one | 1 | 0.02 | 0.04 |
| 1206 | GenJointDISTBOWx200x20one | 1 | 0.01 | 0.02 |
| 966 | GenJointDISTBOWx20x20one | 1 | 0.01 | 0.02 |
| 721 | GenJointNORM5x200x20one | 1 | 0.01 | 0.02 |
| 481 | GenJointNORM5x20x20one | 1 | 0.01 | 0.02 |
| 1201 | GenJointNORMBOWx200x20one | 1 | 0.01 | 0.02 |
| 961 | GenJointNORMBOWx20x20one | 1 | 0.01 | 0.02 |
| 266 | GenSG2x200x20one | 1 | 0.01 | 0.02 |
| 326 | GenSG2x200x50one | 1 | 0.02 | 0.04 |
| 866 | GenSG5x200x100one | 1 | 0.02 | 0.04 |
| 746 | GenSG5x200x20one | 1 | 0.02 | 0.04 |
| 806 | GenSG5x200x50one | 1 | 0.02 | 0.04 |
| 506 | GenSG5x20x20one | 1 | 0.01 | 0.02 |
| 566 | GenSG5x20x50one | 1 | 0.01 | 0.02 |
| 174 | SpecCBOW2x20x100r.two | 1 | 0.01 | 0.02 |
| 234 | SpecCBOW2x20x200r.two | 1 | 0.01 | 0.02 |
| 114 | SpecCBOW2x20x50r.two | 1 | 0.01 | 0.02 |
| 1372 | SpecCBOWBOWx200x100two | 1 | 0.01 | 0.02 |
| 1251 | SpecCBOWBOWx200x20one | 1 | 0.01 | 0.02 |
| 50 | SpecIndyDIST2x20x20dep | 1 | 0.01 | 0.02 |
| 1250 | SpecIndyDISTBOWx200x20dep | 1 | 0.01 | 0.02 |
| 1010 | SpecIndyDISTBOWx20x20dep | 1 | 0.01 | 0.02 |
| 1070 | SpecIndyDISTBOWx20x50dep | 1 | 0.01 | 0.02 |
| 282 | SpecIndyNORM2x200x20two | 1 | 0.01 | 0.02 |
| 762 | SpecIndyNORM5x200x20two | 1 | 0.01 | 0.02 |
| 1238 | SpecJointDISTBOWx200x20r.one | 1 | 0.01 | 0.02 |
| 998 | SpecJointDISTBOWx20x20r.one | 1 | 0.01 | 0.02 |
| 656 | SpecSG5x20x100one | 1 | 0.01 | 0.02 |

kable(top\_r)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | model | precision | recall | Fscore |
| 436 | GenIndyDIST2x200x200one | 0.08 | 0.97 | 0.15 |
| 466 | SpecIndyDIST2x200x200one | 0.08 | 0.97 | 0.15 |
| 468 | SpecIndyDIST2x200x200r.one | 0.08 | 0.97 | 0.15 |
| 946 | SpecIndyDIST5x200x200one | 0.08 | 0.97 | 0.15 |
| 948 | SpecIndyDIST5x200x200r.one | 0.08 | 0.97 | 0.15 |

kable(top\_f)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | model | precision | recall | Fscore |
| 1397 | GenIndyDISTBOWx200x200two | 0.19 | 0.3 | 0.23 |

Visualization of best models:

png(filename = paste(dir, "Best\_parameters.png", sep = "/"))  
plot(eval$precision, eval$recall, xlab = "Precision", ylab = "Recall", main = "Best parameter combination", col = "gray")  
for (i in 1:5){  
 text(top\_p[i,]$precision-0.2, i\*0.05, labels = top\_p[i,]$model, col = "red")  
}  
for (i in 1:5){  
 text(0.2+i\*0.05, top\_r[i,]$recall-i\*0.05, labels = top\_r[i,]$model, col = "blue")  
}  
for (i in 1:5){  
 text(0.2+i\*0.05, top\_f[i,]$Fscore-i\*0.05, labels = top\_f[i,]$model, col = "green")  
}  
dev.off()

## quartz\_off\_screen   
## 2