HW2

Dan Zalewski

19Aug15

## Flights at ABIA

### Required Files:

* ABIA.csv ###### Which airlines cancel the most flights?

ggplot(ABIA, aes(UniqueCarrier, Cancelled)) + geom\_bar(stat = "identity") + labs(title = "Flight Cancellations by Airline", x = "Airline", y = "Cancellations")

## Author Attribution

### Required Files:

* c50train folder (50 .txt files)
* c50test folder (50 .txt files)

### Model 1: Naive Bayes

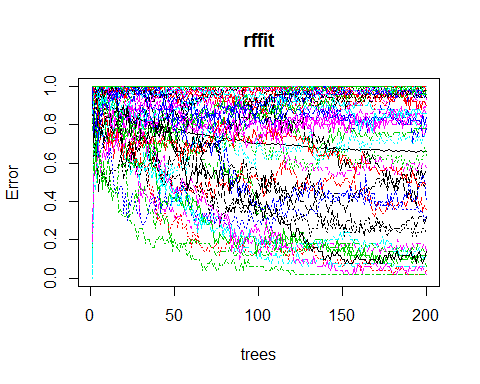
## Loading required package: NLP

##### The Naive Bayes model correctly matched the authors to their works 0.9732% of the time.

### Model 2: Random Forest

## randomForest 4.6-10  
## Type rfNews() to see new features/changes/bug fixes.

#### Random Forest Visualization



## [1] 18

## [1] 2500

## AaronPressman AlanCrosby AlexanderSmith BenjaminKangLim   
## 4 119 4 95   
## BernardHickey BradDorfman DarrenSchuettler DavidLawder   
## 55 6 85 68   
## EdnaFernandes EricAuchard FumikoFujisaki GrahamEarnshaw   
## 91 5 50 48   
## HeatherScoffield JaneMacartney JanLopatka JimGilchrist   
## 33 62 31 73   
## JoeOrtiz JohnMastrini JonathanBirt JoWinterbottom   
## 22 80 4 227   
## KarlPenhaul KeithWeir KevinDrawbaugh KevinMorrison   
## 40 112 35 84   
## KirstinRidley KouroshKarimkhany LydiaZajc LynneO'Donnell   
## 6 248 39 66   
## LynnleyBrowning MarcelMichelson MarkBendeich MartinWolk   
## 53 83 6 1   
## MatthewBunce MichaelConnor MureDickie NickLouth   
## 52 10 60 99   
## PatriciaCommins PeterHumphrey PierreTran RobinSidel   
## 22 122 4 13   
## RogerFillion SamuelPerry SarahDavison ScottHillis   
## 34 26 0 34   
## SimonCowell TanEeLyn TheresePoletti TimFarrand   
## 11 5 8 20   
## ToddNissen WilliamKazer   
## 44 1

##### The random forest model correctly matched the authors to their works 0.16% of the time.

## Association Rule Mining

### Required Files:

* groceries.txt

#### A priori algorithm:

## Loading required package: Matrix  
##   
## Attaching package: 'arules'  
##   
## The following object is masked from 'package:tm':  
##   
## inspect  
##   
## The following objects are masked from 'package:base':  
##   
## %in%, write

##   
## Parameter specification:  
## confidence minval smax arem aval originalSupport support minlen maxlen  
## 0.5 0.1 1 none FALSE TRUE 0.005 1 10  
## target ext  
## rules FALSE  
##   
## Algorithmic control:  
## filter tree heap memopt load sort verbose  
## 0.1 TRUE TRUE FALSE TRUE 2 TRUE  
##   
## apriori - find association rules with the apriori algorithm  
## version 4.21 (2004.05.09) (c) 1996-2004 Christian Borgelt  
## set item appearances ...[0 item(s)] done [0.00s].  
## set transactions ...[20509 item(s), 40075 transaction(s)] done [0.02s].  
## sorting and recoding items ... [48 item(s)] done [0.00s].  
## creating transaction tree ... done [0.01s].  
## checking subsets of size 1 2 3 done [0.00s].  
## writing ... [1 rule(s)] done [0.00s].  
## creating S4 object ... done [0.00s].

inspect(groceriesrules)

## lhs rhs support confidence lift  
## 1 {curd} => {whole milk} 0.005190268 0.5 11.67686

inspect(subset(groceriesrules, subset=lift > 3))

## lhs rhs support confidence lift  
## 1 {curd} => {whole milk} 0.005190268 0.5 11.67686

inspect(subset(groceriesrules, subset=confidence > 0.65))

## NULL

inspect(subset(groceriesrules, subset=support > .005 & confidence > 0.65))

## NULL

#### Interpretation:

Confidence of 0.5 means that at least half of the time the first item is purchased, the second one will also be purchased.