DMC@ISU: Iowa State University Data Mining Cup Team 2015

Initial Exploration

Spring 2015, Iowa State

Due Date: April 22 2015

I am using the following packages:

```
library(ggplot2)
library(lubridate)
library(dplyr)
library(reshape2)
library(sqldf)
```

and my working directory is set to `dmc2015/ian`.

0.1 Reading the Data

Read the (clean) data into R:

```
# training set ('historical data')
trn.raw <- read.delim("~/dmc2015/data/raw_data/DMC_2015_orders_train.txt",
    stringsAsFactors = FALSE, sep = "|", quote = "")
trn <- read.csv("~/dmc2015/data/clean_data/train_simple_name.csv",
    stringsAsFactors = FALSE, na.strings = "")[, names(trn.raw)]

# test set ('future data')
cls.raw <- read.delim("~/dmc2015/data/raw_data/DMC_2015_orders_class.txt",
    stringsAsFactors = FALSE, sep = "|", quote = "")
cls <- read.csv("~/dmc2015/data/clean_data/test_simple_name.csv",
    stringsAsFactors = FALSE, na.strings = "")[, names(cls.raw)]
```

0.2 Reading the Features

Read the batch features:

```
# batch features in
# ~/dmc2015/features/feature_files/batchInfo_test.csv,
# batchInfo_train.csv
batchInfo_train <- readRDS("../batchInfo_train.rds")
batchInfo_test <- readRDS("../batchInfo_test.rds")
```

Add the batch features:

```
trn <- trn %>% left_join(batchInfo_train, by = "orderID")
cls <- cls %>% left_join(batchInfo_test, by = "orderID")
```

0.3 Writing the Feature Matrix

Save the results as a list

```
write.csv(trn, file = "~/dmc2015/data/featureMatrix/train_ver0.0.csv",
    row.names = FALSE, na = "", quote = FALSE)
write.csv(cls, file = "~/dmc2015/data/featureMatrix/class_ver0.0.csv",
    row.names = FALSE, na = "", quote = FALSE)

featMat <- list(train = trn, class = cls)
saveRDS(featMat, file = "~/dmc2015/data/featureMatrix/featMat_v0.0.rds")
```