Home Depot Term Project

Midterm Status Report

Chunyao Wang,

Wenxing Li,

Jie Rong,

Yifan Zhao,

Nanxun Xie

1. What is the Kaggle competition?

Kaggle competition is a predictive modeling and analytics competition which data miners from all over the world compete to produce the best models for the data and statisticians which are posted by companies and researchers. Because there is a wide variety of strategies that can be used to any predictive modelling task and it is impossible for the company or researchers to know which kind of technique or analyst is most effective, the Kaggle competition is a very useful way to find the most satisfying approach. Kaggle competitions have left great impact on all kinds of fields including HIV research, chess ratings and traffic forecasting.

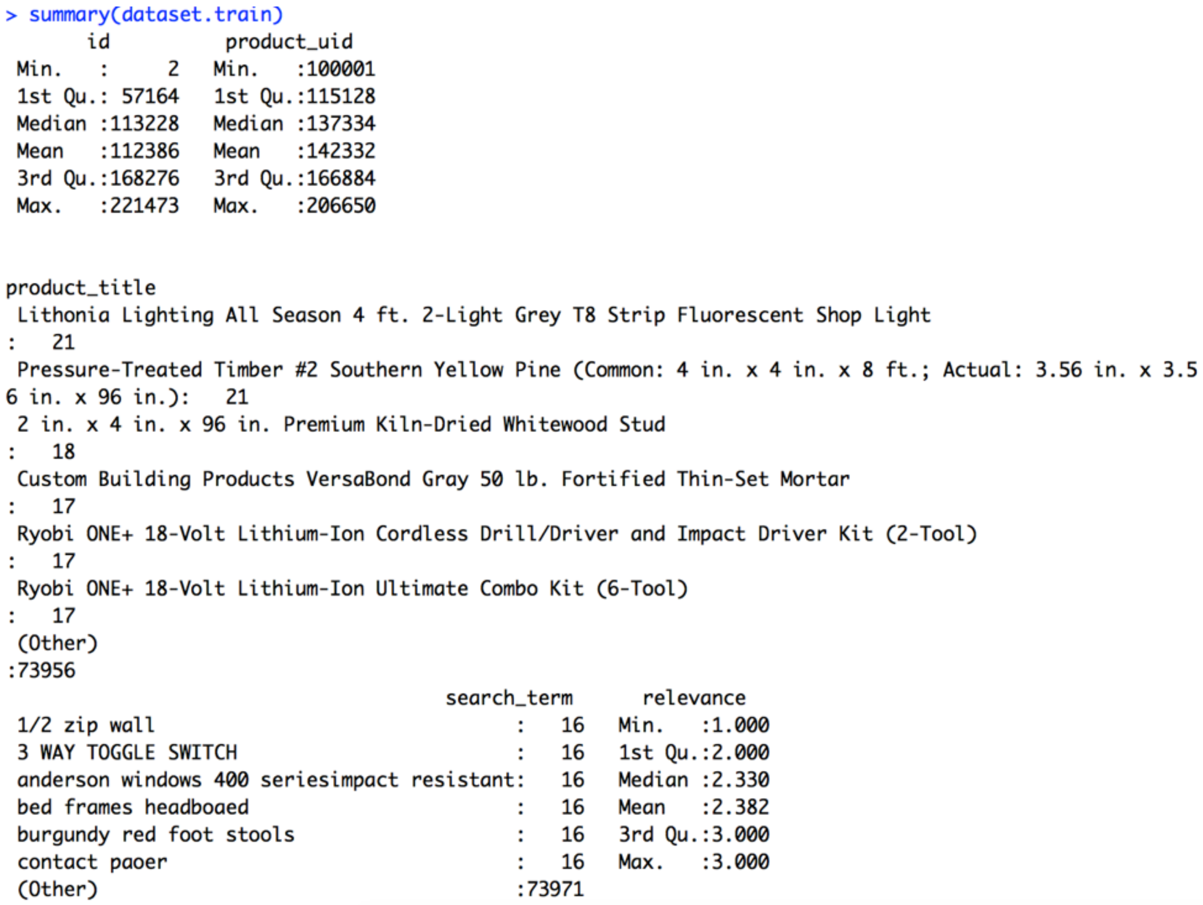
2. What have we done?

This home competition is in order to find out the extent and relevance of a search result matches the search query which is paired with. All our work is based upon the datasets provided then further to judge. First we inserted and analysized the train.csv, and by checking whether there is any missing value to check the datasets, and by reviewing the first 6 data to create the plot to illustrate the changes of distance according to the rate.

Following are our R file and some explanations:

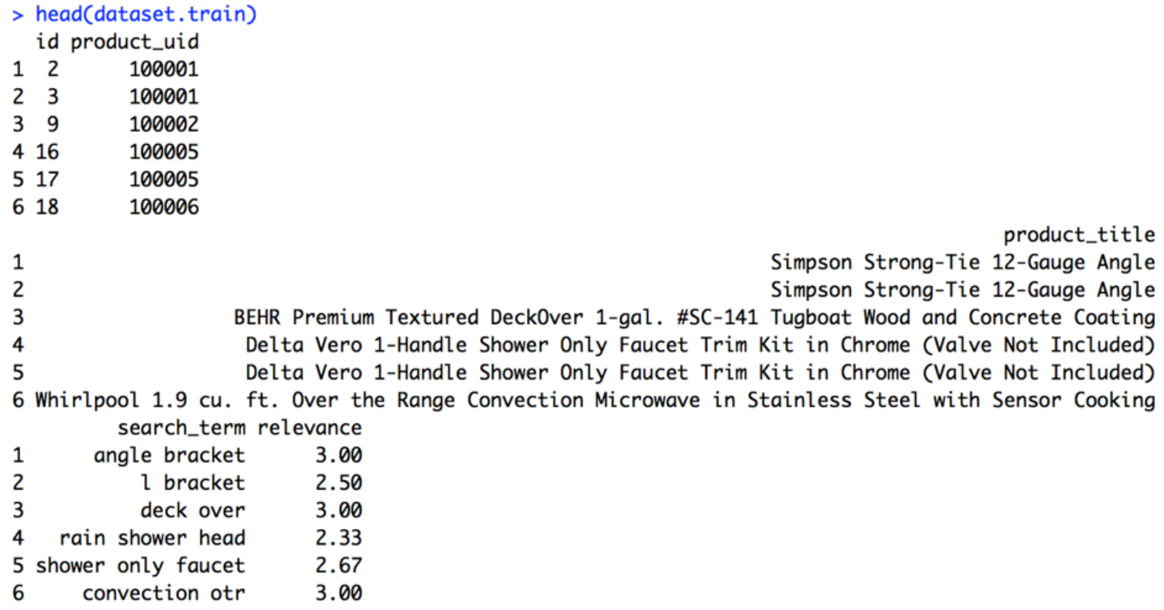
```
rm(list = ls())
dataset.train<-read.csv("train.csv") ##read train.csv file
dataset.test<-read.csv("test.csv") ##read test.csv file
```

```
summary(dataset.train) ##general analysis of train.csv
```



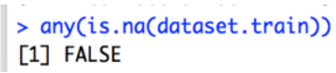
```
> summary(dataset.train)
       id           product_uid    
 Min.   :     2   Min.   :100001  
 1st Qu.: 57164   1st Qu.:115128  
 Median :113228   Median :137334  
 Mean   :112386   Mean   :142332  
 3rd Qu.:168276   3rd Qu.:166884  
 Max.   :221473   Max.   :206650  


product_title                                                                                     
 Lithonia Lighting All Season 4 ft. 2-Light Grey T8 Strip Fluorescent Shop Light                  
:   21  
 Pressure-Treated Timber #2 Southern Yellow Pine (Common: 4 in. x 4 in. x 8 ft.; Actual: 3.56 in. x 3.5
6 in. x 96 in.):   21  
 2 in. x 4 in. x 96 in. Premium Kiln-Dried Whitewood Stud                                         
:   18  
 Custom Building Products VersaBond Gray 50 lb. Fortified Thin-Set Mortar                         
:   17  
 Ryobi ONE+ 18-Volt Lithium-Ion Cordless Drill/Driver and Impact Driver Kit (2-Tool)              
:   17  
 Ryobi ONE+ 18-Volt Lithium-Ion Ultimate Combo Kit (6-Tool)                                       
:   17  
 (Other)                                                                                          
:73956  
                                     search_term      relevance    
 1/2 zip wall                              :   16   Min.   :1.000  
 3 WAY TOGGLE SWITCH                       :   16   1st Qu.:2.000  
 anderson windows 400 seriesimpact resistant:   16   Median :2.330  
 bed frames headboaed                      :   16   Mean   :2.382  
 burgundy red foot stools                  :   16   3rd Qu.:3.000  
 contact paoer                             :   16   Max.   :3.000  
 (Other)                                   :73971                  
```

head(dataset.train) **## review first six elements in train.csv**



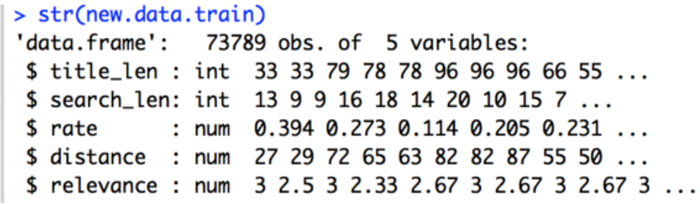
```
> head(dataset.train)
  id product_uid
1  2      100001
2  3      100001
3  9      100002
4 16      100005
5 17      100005
6 18      100006
                                                                                product_title
1                                                            Simpson Strong-Tie 12-Gauge Angle
2                                                            Simpson Strong-Tie 12-Gauge Angle
3                        BEHR Premium Textured DeckOver 1-gal. #SC-141 Tugboat Wood and Concrete Coating
4                         Delta Vero 1-Handle Shower Only Faucet Trim Kit in Chrome (Valve Not Included)
5                         Delta Vero 1-Handle Shower Only Faucet Trim Kit in Chrome (Valve Not Included)
6 Whirlpool 1.9 cu. ft. Over the Range Convection Microwave in Stainless Steel with Sensor Cooking
         search_term relevance
1      angle bracket      3.00
2          l bracket      2.50
3          deck over      3.00
4   rain shower head      2.33
5 shower only faucet      2.67
6     convection otr      3.00
```

```
any(is.na(dataset.train))##does it contain missing value
```



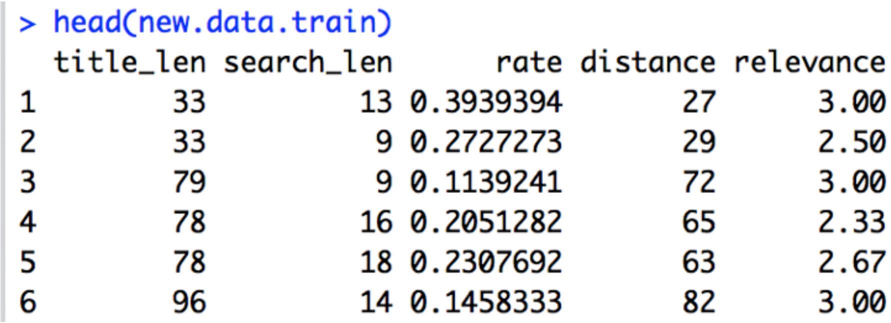
```
> any(is.na(dataset.train))
[1] FALSE
```

```
library("stringdist")
dataset.train$product_title<-as.character(dataset.train$product_title)
dataset.train$search_term<-as.character(dataset.train$search_term)
distance<-stringdist(dataset.train$product_title,dataset.train$search_term)
title_len<-nchar(dataset.train$product_title,allowNA = T)
search_len<-nchar(dataset.train$search_term,allowNA = T)
rate=search_len/title_len

new.data.train=data.frame(title_len=title_len,search_len=search_len,rate=rate,distance=distance,relevance=dataset.train$relevance)
remove=which(is.na(new.data.train))
new.data.train<-new.data.train[-remove,]##remove missing values

str(new.data.train)##show data frame of new.data.train
```



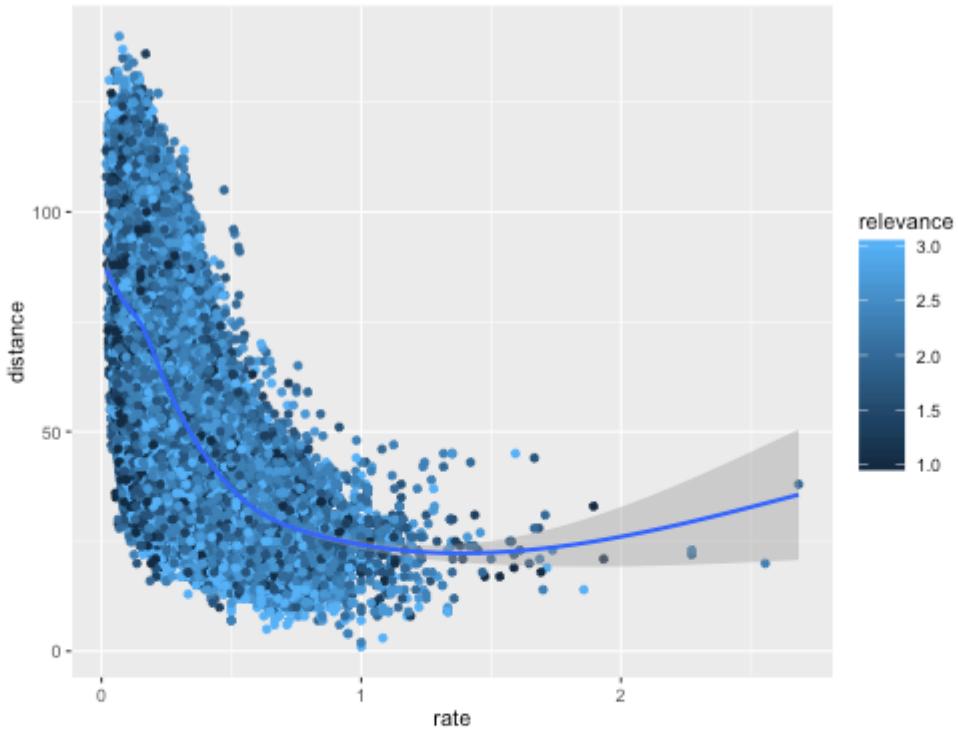
```
> str(new.data.train)
'data.frame':	73789 obs. of  5 variables:
 $ title_len : int  33 33 79 78 78 96 96 96 66 55 ...
 $ search_len: int  13 9 9 16 18 14 20 10 15 7 ...
 $ rate      : num  0.394 0.273 0.114 0.205 0.231 ...
 $ distance  : num  27 29 72 65 63 82 82 87 55 50 ...
 $ relevance : num  3 2.5 3 2.33 2.67 3 2.67 3 2.67 3 ...
```

```
head(new.data.train)
```



```
> head(new.data.train)
  title_len search_len      rate distance relevance
1        33         13 0.3939394       27      3.00
2        33          9 0.2727273       29      2.50
3        79          9 0.1139241       72      3.00
4        78         16 0.2051282       65      2.33
5        78         18 0.2307692       63      2.67
6        96         14 0.1458333       82      3.00
```

```
library(ggplot2ggplot(new.data.train,aes(x=rate,y=distance,col=relevance))+geom_point()+geom_smooth() ##plot the changes of distance according to rate
```



3. What do we plan to do in the future?

Our main goal is to build a system which can simulate the relevant scores of search results. As we already figured out the basic statistical features of data train.csv, we need to figure out the relationships between search term, production title and relevant score next. The main principle we use is string matching and we decide to use stringdist package in Rstudio to process it. We also plan to split strings in production title and search terms and classify them for further use.