**Home Depot Term Project**

**Midterm Status Report**

**Chunyao Wang,**

**Wenxing Li,**

**Jie Rong,**

**Yifan Zhao,**

**Nanxun Xie**

**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.

**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.

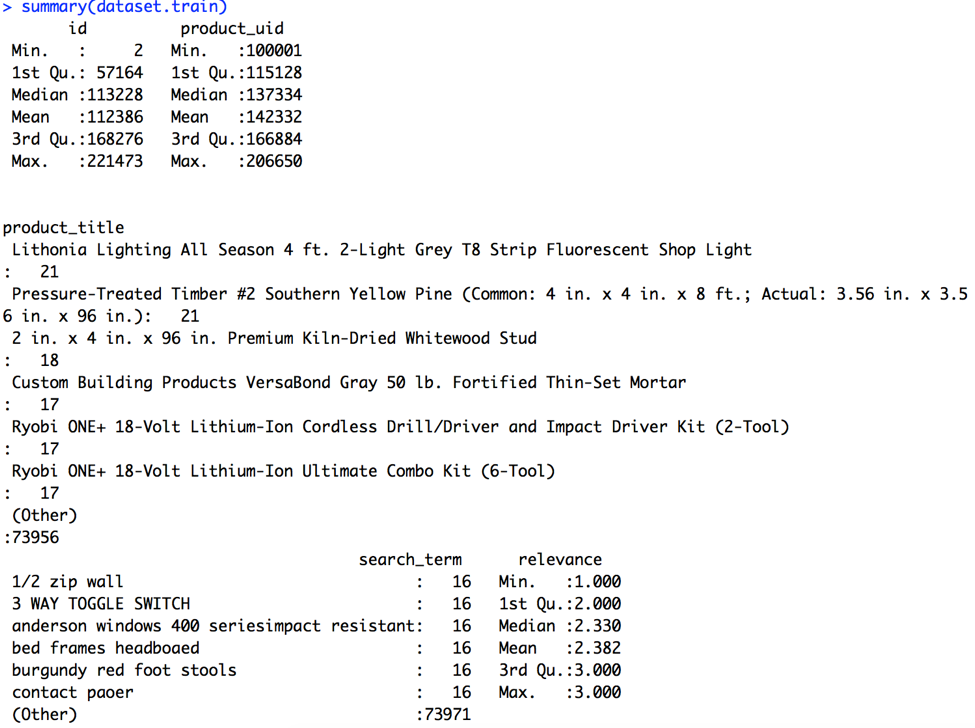
**Following are our R file and some explanations:**

rm(list = ls())

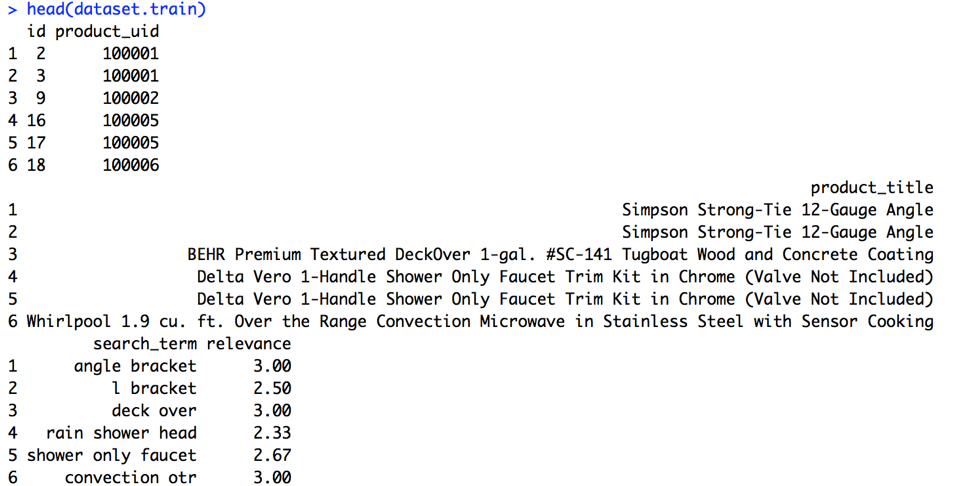
dataset.train<-read.csv("train.csv") ##read train.csv file

dataset.test<-read.csv("test.csv") ##read test.csv file

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



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



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

Picture3.png

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)

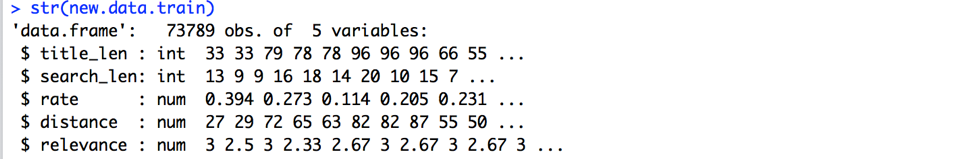
rate=search\_len/title\_len

new.data.train=data.frame(title\_len=title\_len,search\_len=search\_len,rate=rate,distance=distance,relevance=dataset.train$relevance)

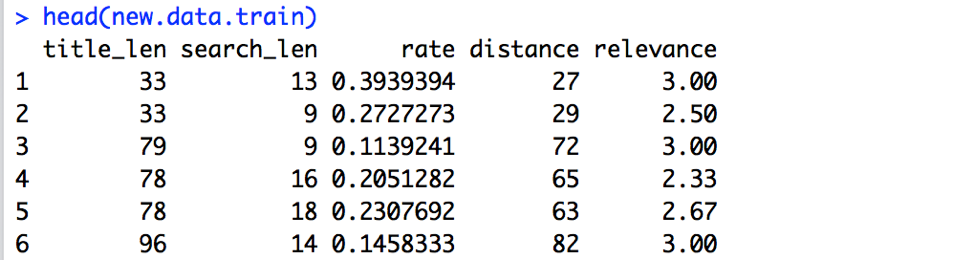
remove=which(is.na(new.data.train))

new.data.train<-new.data.train[-remove,]##remove missing values

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

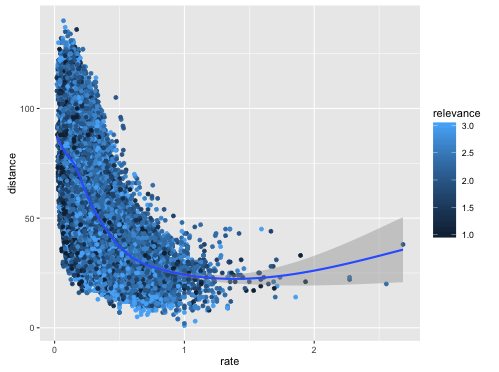


head(new.data.train)



library(ggplot2

ggplot(new.data.train,aes(x=rate,y=distance,col=relevance))+geom\_point()+geom\_smooth()##plot the changes of distance according to rate



**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 analysed the basic statistical features of data train.csv, we need to figure out the relationships bewteen 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.