Project ML Fall 2015. Appendix 1

Anastasiya Yarygina Udovenko, Manuel Aragonés Mora, Andrés Ponce de Leon

December 7, 2015

```
rm(list=ls())

setwd("~/Dropbox/MPP/ML/project")
```

I Similarity analysis

```
# load packages
```

```
library(recommenderlab)
library(ggplot2)
```

```
# read data
data<- read.csv("dataNA.csv", sep="," , header=TRUE)
dim(data)
```

```
## [1] 3146  630
```

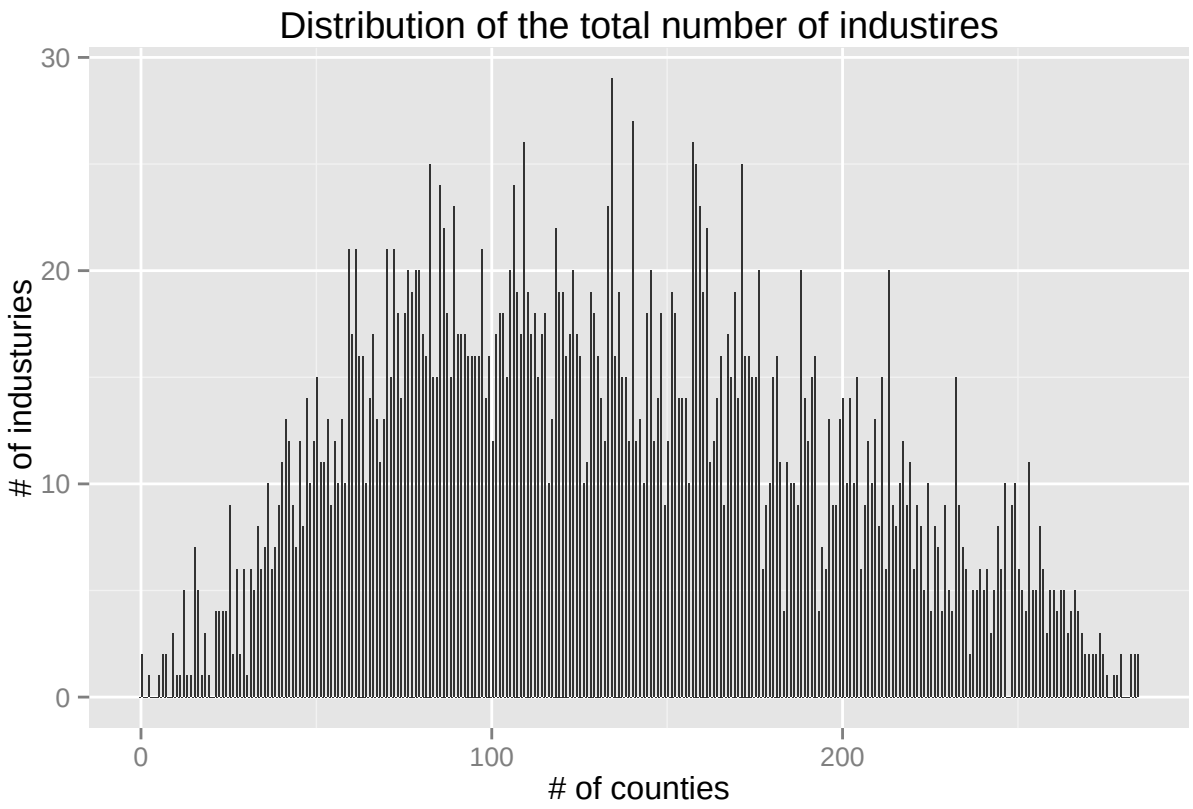
```
# remove last three rows:
n<-dim(data)[1]
data<-data[1:(n-3),]

# get rca matrix
rcaData<- data[,grep("rca", colnames(data))]
# get county ids
countyName <- data$NAME
stateName <- data$STATE_NAME
ID<- with(data, paste0(NAME,", ", STATE_NAME))
rownames(rcaData) <- ID

# convert into matrix:
matrix<- data.matrix(rcaData)
rrm = as(matrix, "realRatingMatrix")
```

Preliminar data exploration

```
#Some summary statistics of the distributino of the rankings:
# average index of competitiveness
summary(getRatings(rrm))[4]
```

```
##   Mean
## 0.5801
```

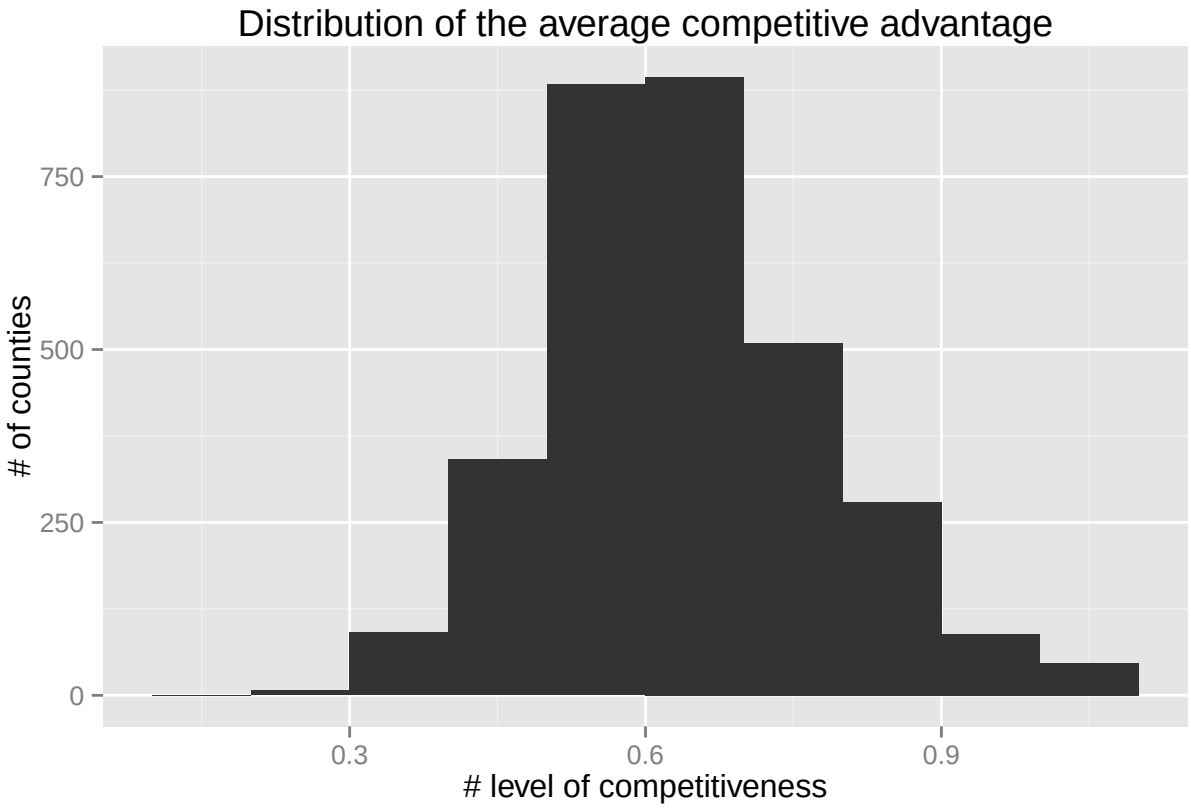
```
# How many industries do counties have?
qplot(rowCounts(rrm), binwidth = 0.5,
      main = "Distribution of the total number of industires",
      xlab = "# of counties",
      ylab = "# of industuries")
```




```
# the average county has :
summary(rowCounts(rrm))[4]
```

```
##  Mean
## 133.1
```

```
# industries

# What is the average level of competitive advantage in counties across industires?
qplot(rowMeans(rrm), binwidth = 0.1,
      main = "Distribution of the average competitive advantage",
      xlab = "# level of competitiveness",
      ylab = "# of counties")
```

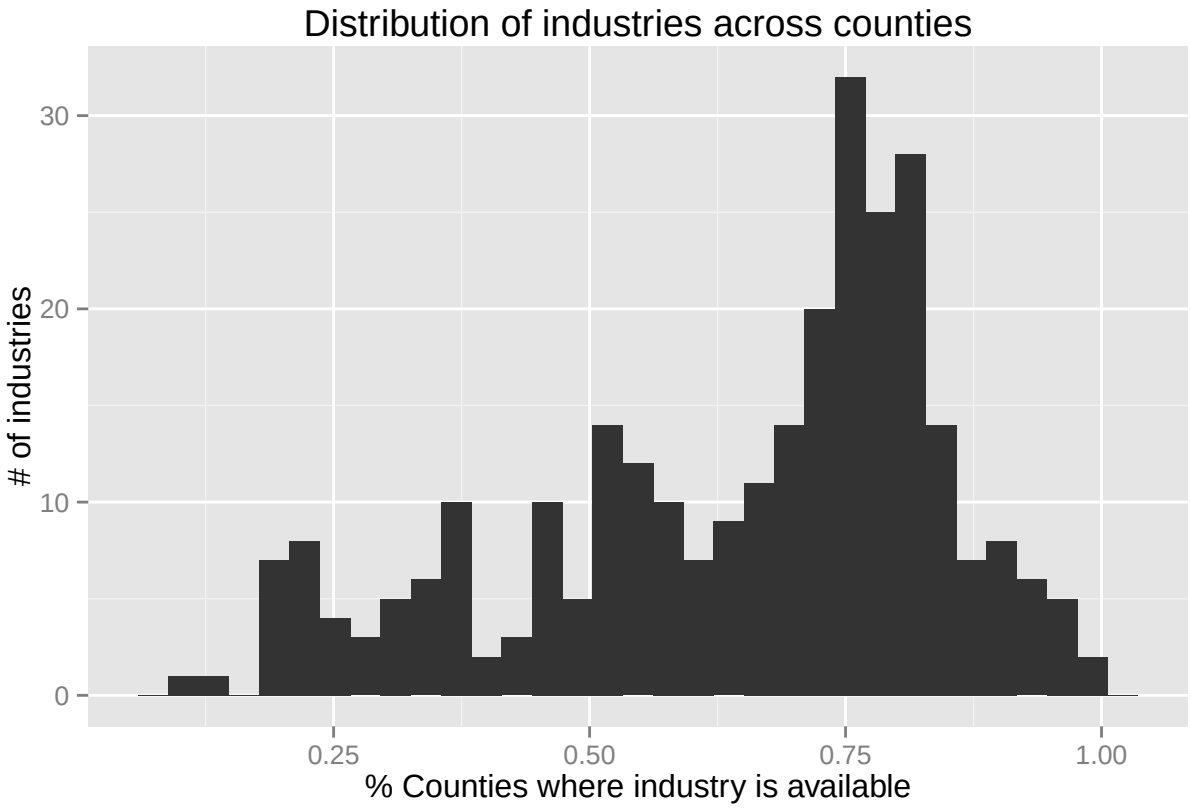



```
# the average level of competitive advantage is:
summary(rowMeans(rrm))[4]
```

```
##  Mean
## 0.638
```

```
# What is the average importance of industries (share of counties that have given industry)?
qplot(colMeans(rrm),
    main = "Distribution of industries across counties",
    xlab = "% Counties where industry is available",
    ylab = "# of industries")
```

```
## stat_bin: binwidth defaulted to range/30. Use 'binwidth = x' to adjust this.
```



some industries are availalble in all counties some only in few. On average:

```
summary(colMeans(rrm))[4]
```

```
##   Mean
## 0.6463
```

Find counties that have the largest number of industires

```
# coerce data into dataframe:
df<- (as(rrm, "data.frame"))
# Sort the dataframe, order in decreasing order, find counties with most number of industries.
sort(table(df$user),decreasing=TRUE)[1:3]
```

```
##
##          Harris, Texas Los Angeles, California         Cook, Illinois
##                    284                     284                    283
```

Indutires present in largest number of counties:

```
sort(table(df$item),decreasing=TRUE)[1:3]
```

```
##
## rca_4471 rca_7225 rca_5221
##     3118     3118     3114
```

```r
b<-sort(table(df$item),decreasing=TRUE)[1:10]

# codes of the most popular industres:
# strip "rca" part and get only numbers
s1 = unlist(strsplit(names(b)[1], split='_', fixed=TRUE))[2]
s2 = unlist(strsplit(names(b)[2], split='_', fixed=TRUE))[2]
s3 = unlist(strsplit(names(b)[3], split='_', fixed=TRUE))[2]
s4 = unlist(strsplit(names(b)[4], split='_', fixed=TRUE))[2]

# load the table with industries identifiers
industries<- read.csv("industries1.csv", sep="," , header=TRUE)
head(industries)
```

```
##   code                                          name
## 1 1111                   Oilseed and Grain Farming
## 2 1112                 Vegetable and Melon Farming
## 3 1113                  Fruit and Tree Nut Farming
## 4 1114 Greenhouse, Nursery, and Floriculture Production
## 5 1119                           Other Crop Farming
## 6 1121                 Cattle Ranching and Farming
```

```r
# most popular industires are:
industries[ which(industries$code==s1), ]
```

```
##     code              name
## 155 4471 Gasoline Stations
```

```r
industries[ which(industries$code==s2), ]
```

```
##     code                               name
## 290 7225 Restaurants and Other Eating Places
```

```r
industries[ which(industries$code==s3), ]
```

```
##     code                             name
## 212 5221 Depository Credit Intermediation
```

```r
industries[ which(industries$code==s4), ]
```

```
##     code                     name
## 299 8131 Religious Organizations
```

```r
#get counties with the highest average index of competitveness.
order_simUser = order(rowMeans(rrm), decreasing = F)
rownames(rrm)[order_simUser[1:10]]
```

```
##  [1] "Fairfax County, Virginia" "Bronx, New York"
##  [3] "Queens, New York"         "Santa Clara, California"
##  [5] "Nassau, New York"         "Cape May, New Jersey"
##  [7] "Collin, Texas"            "Palm Beach, Florida"
##  [9] "Montgomery, Maryland"     "Lee, Florida"
```

Find similar counties:

```
#normalize the data
rdn = normalize(rrm)

# find counties similar to the first county in the list
sim = similarity(x=rdn[1,], y=rdn[-1,], method="cosine")
order_sim = order(sim, decreasing = T)

# these are ten most similar counties:
order_sim[1:10]
```

```
##  [1] 2481 2683 2181 2467 1073 1455 2643  785  662 1234
```

```
# these are their normalized level of competitiveness:
sim[order_sim[1:10]]
```

```
##  [1] 0.4370291 0.3924489 0.3904927 0.3873222 0.3860016 0.3769312 0.3766376
##  [8] 0.3763083 0.3745861 0.3743404
```

```
#  This county
rownames(rdn)[1]
```

```
## [1] "Autauga, Alabama"
```

```
# is similar to these counties:
rownames(rdn)[order_sim[1:10]]
```

```
##  [1] "Loudon, Tennessee"    "Mason, Texas"         "Murray, Oklahoma"
##  [4] "Henderson, Tennessee" "Martin, Kentucky"     "Panola, Mississippi"
##  [7] "Jackson, Texas"       "Washington, Indiana"  "Monroe, Illinois"
## [10] "Allegan, Michigan"
```

Find countieis similar to Cook County, Illinois.

```
# get cookcounty id
Cookcounty<- rownames(rdn)[grep("Cook, Illinois", rownames(rcaData))]

# get the id of Cookcounty in ths matrix:
county <- c(Cookcounty)
countyID<- match(county, rownames(rdn))
# the County's id is:
countyID
```

```
## [1] 611
```

```
# find similar counties
sim1 = similarity(x=rdn[countyID,], y=rdn[-countyID,], method="cosine")
order_sim1 = order(sim1, decreasing = T)
# these are ten most similar counties:
order_sim1[1:10]
```

```
##  [1]  205  616 2060 1775 1340 1293  224 1857  447 1229
```

```
# these are their normlized levels of competitive advantage:
sim[order_sim1[1:10]]
```

```
##  [1]  0.13309152 -0.23442615 -0.16016471 -0.27780145 -0.19739886
##  [6] -0.20179663 -0.04317568 -0.25994487  0.18685844 -0.12010312
```

```
#  This county:
rownames(rrm)[countyID]
```

```
## [1] "Cook, Illinois"
```

```
# is similar  counties:
rownames(rrm)[order_sim1[1:10]]
```

```
##  [1] "Los Angeles, California"   "Douglas, Illinois"
##  [3] "Crawford, Ohio"            "Atlantic, New Jersey"
##  [5] "Grant, Minnesota"          "Newaygo, Michigan"
##  [7] "San Francisco, California" "Montgomery, New York"
##  [9] "Fulton, Georgia"           "Plymouth, Massachusetts"
```

Make Recommendations to Lake County, Indiana

```
LakeCounty<- rownames(rrm)[grep("Lake, Indiana", rownames(rrm))]

# get the id of Lake County in ths matrix:
county <- c(LakeCounty)
countyID<- match(county, rownames(rrm))
# the County's id is:
countyID
```

```
## [1] 742
```

Recommend industires to Lake County:

```
r1 <- Recommender(rrm, method = "POPULAR")

# b. Recommend 5 industries to Lake County
recom1 <- predict(r1, rrm[countyID], n=5, type="topNList")

# get top 5 industries from the list of recommendations
as(recom1, "list")
```

```
## [[1]]
## [1] "rca_3211" "rca_7121" "rca_7212" "rca_1133" "rca_1153"
```

```
# recommended industries are:
industries[ which(industries$code=="3211"), ]
```

```
##    code                          name
## 60 3211 Sawmills and Wood Preservation
```

```
industries[ which(industries$code=="7121"), ]
```

```
##     code                                                name
## 281 7121 Museums, Historical Sites, and Similar Institutions
```

```
industries[ which(industries$code=="7212"), ]
```

```
##     code                                                  name
## 286 7212 RV (Recreational Vehicle) Parks and Recreational Camps
```

```
industries[ which(industries$code=="1133"), ]
```

```
##    code    name
## 14 1133 Logging
```

```
industries[ which(industries$code=="1153"), ]
```

```
##    code                          name
## 19 1153 Support Activities for Forestry
```

Evaluation of recommender method

```
# set the evaluation shceme

# focus only on popular industries that:
# 1) are availalbe in at least 3 counties
rrm = rrm[,colCounts(rrm) > 3]

# 2) we keep counties that have at least 2 industries
rrm = rrm[rowCounts(rrm) > 2,]

#set scheme
scheme <- evaluationScheme(rrm, method = "split", train = .9,
                           k = 1, given = 2, goodRating = 1)

# create the list of methods (algotims) to analyze
algorithms <- list(
    "random method" = list(name="RANDOM", param=list(normalize = "Z-score")),
    "popular method" = list(name="POPULAR", param=list(normalize = "Z-score")),
    "user-based CF" = list(name="UBCF", param=list(normalize = "Z-score",
                                                   method="Cosine",
                                                   nn=50, minRating=1)))
```

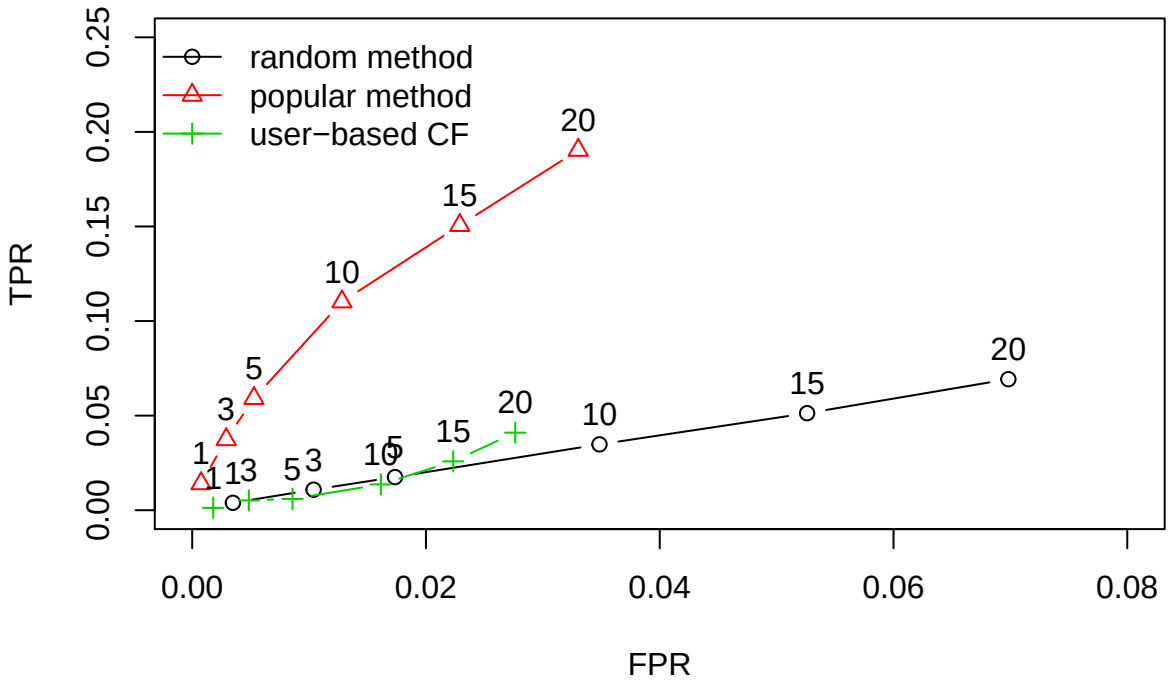
```
# run algorithms and predict next n industires
results <- evaluate(scheme, algorithms, n=c(1, 3, 5, 10, 15, 20))
```

```
## RANDOM run
##   1  [0.005sec/0.293sec]
## POPULAR run
##   1  [0.163sec/0.146sec]
## UBCF run
##   1  [0.172sec/2.981sec]
```

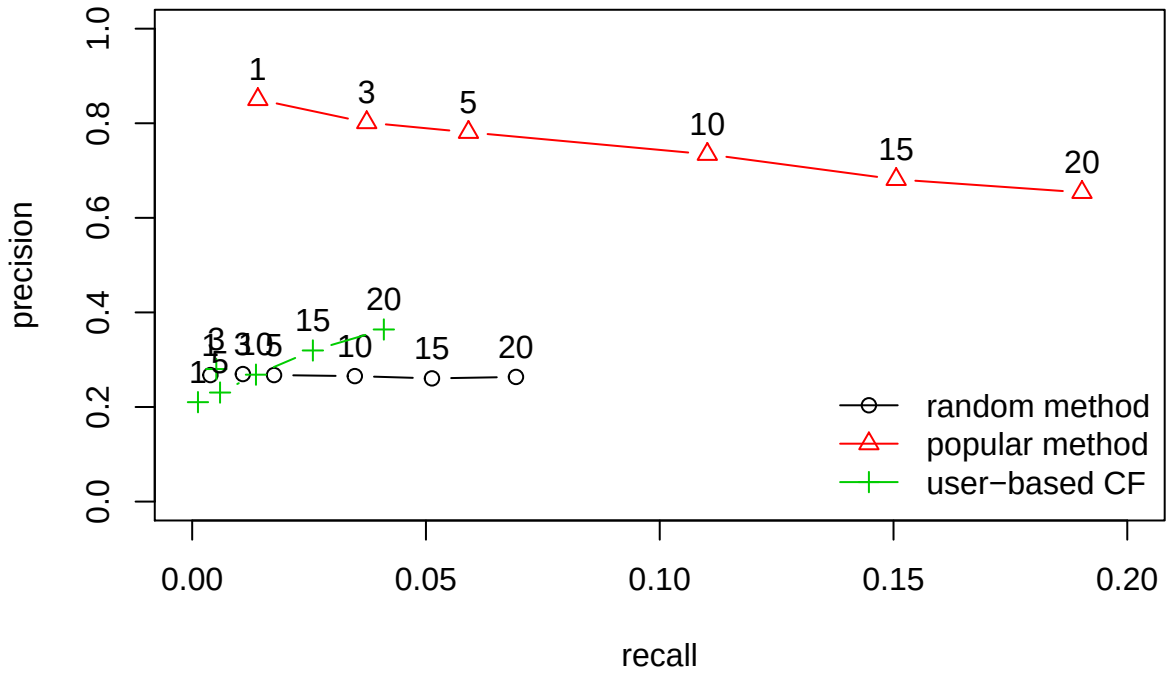
```
# compare predicted ratings for several methods
evlist <- evaluate(scheme, algorithms, type="ratings")
```

```
## RANDOM run
##   1  [0.005sec/0.028sec]
## POPULAR run
##   1  [0.159sec/0.03sec]
## UBCF run
##   1  [0.151sec/2.84sec]
```

```
# compare ROC curves for the selected evaluation scheme.
```




```
# Note: TPR is true positive rate and FPR is False positive rate.

# Comparison of precision-recall curves for several recommender methods
```


precision
recall
0.0
0.2
0.4
0.6
0.8
1.0
0.00
0.05
0.10
0.15
0.20
random method
popular method
user−based CF