# Recommendation by cosine similarity

Krishanu Banerjee

Tuesday, September 27, 2016

Objective is to find similarity between institutions so that application can recommend which insitutions are the best fit for a particular student. Students will provide his/her background information and application will find the best fit institution for the student based on cosine similarity.

Data preperation- Attributes are based on mainly students background, family background, interest of study and cost of study. The name of the institution and zip code combination are considered as unique key.Cost for the program is selected based on private, public or others. Multiple years of data averaged over unique key.

```
library(RSQLite)

## Warning: package 'RSQLite' was built under R version 3.1.3

## Loading required package: DBI

## Warning: package 'DBI' was built under R version 3.1.3

library('magrittr')

## Warning: package 'magrittr' was built under R version 3.1.3

library('tidyr')

## Warning: package 'tidyr' was built under R version 3.1.3

## 
## Attaching package: 'tidyr'
## 
## The following object is masked from 'package:magrittr':
## 
##     extract

library('dplyr')

## Warning: package 'dplyr' was built under R version 3.1.3

## 
## Attaching package: 'dplyr'
## 
## The following objects are masked from 'package:stats':
## 
##     filter, lag
## 
## The following objects are masked from 'package:base':
```

```
##
##     intersect, setdiff, setequal, union

library('ggplot2')

## Warning: package 'ggplot2' was built under R version 3.1.3

library('gridExtra')

## Warning: package 'gridExtra' was built under R version 3.1.3

##
## Attaching package: 'gridExtra'
##
## The following object is masked from 'package:dplyr':
##
##     combine

library('bnlearn')

## Warning: package 'bnlearn' was built under R version 3.1.3

library('leaflet')

## Warning: package 'leaflet' was built under R version 3.1.3

library('htmltools')

## Warning: package 'htmltools' was built under R version 3.1.3

library('RColorBrewer')

## Warning: package 'RColorBrewer' was built under R version 3.1.3

library("gplots")

## Warning: package 'gplots' was built under R version 3.1.3

##
## Attaching package: 'gplots'
##
## The following object is masked from 'package:stats':
##
##     lowess

library("cluster")

## Warning: package 'cluster' was built under R version 3.1.3

db <- dbConnect(dbDriver("SQLite"),
"D:/college_score/output/database.sqlite")


trainSummary<- dbGetQuery(db, "
                   select LOWER(INSTNM) || '-' || ZIP AS INSTNM_ZIP
                          ,AVG(SAT_AVG_ALL)
                          ,CASE WHEN COSTT4_A = 'NA' THEN AVG(COSTT4_P) ELSE
AVG(COSTT4_A) END as COSTT --cost anual academic year
                           ,AVG(PCTFLOAN) --federal loan rate
                          ,AVG(UG25abv) -- undergrad over age 25
                          ,AVG(PAR_ED_N)
                          ,AVG(PAR_ED_PCT_1STGEN)
                          ,AVG(DEBT_MDN) -- median debt
            ,AVG(gt_25k_p6) -- shareof students earning over 25k after 6 year
                          ,AVG(NONCOM_RPY_5YR_RT)
                          ,AVG(first_gen)
                          , AVG(md_faminc)
                          ,AVG(pct_ba)
                          , AVG(pct_grad_prof)
                          , AVG(median_hh_inc)
                          , AVG(unemp_rate)
                          ,AVG(loan_ever)
                           ,AVG(PCIP01)
 ,AVG(PCIP03)
 ,AVG(PCIP04)
 ,AVG(PCIP05)
 ,AVG(PCIP09)
 ,AVG(PCIP10)
 ,AVG(PCIP11)
 ,AVG(PCIP12)
 ,AVG(PCIP13)
 ,AVG(PCIP14)
 ,AVG(PCIP15)
 ,AVG(PCIP16)
 ,AVG(PCIP19)
 ,AVG(PCIP22)
 ,AVG(PCIP23)
 ,AVG(PCIP24)
 ,AVG(PCIP25)
 ,AVG(PCIP26)
 ,AVG(PCIP27)
 ,AVG(PCIP29)
 ,AVG(PCIP30)
 ,AVG(PCIP31)
 ,AVG(PCIP38)
 ,AVG(PCIP39)
 ,AVG(PCIP40)
 ,AVG(PCIP41)
 ,AVG(PCIP42)
 ,AVG(PCIP43)
 ,AVG(PCIP44)
 ,AVG(PCIP45)
 ,AVG(PCIP46)
 ,AVG(PCIP47)
 ,AVG(PCIP48)
 ,AVG(PCIP49)
 ,AVG(PCIP50)
 ,AVG(PCIP51)
 ,AVG(PCIP52)
 ,AVG(PCIP54)
,ifnull(case when (NPT4_OTHER is null and NPT4_PRIV is null and NPT4_PROG is
null) then  AVG(NPT4_PUB)
      when (NPT4_OTHER is null and  NPT4_PUB is null and NPT4_PROG is null)
then  AVG(NPT4_PRIV)
      when (NPT4_OTHER is null and  NPT4_PUB is null and NPT4_PRIV is null)
then  AVG(NPT4_PROG)
      when (NPT4_PROG is null and  NPT4_PUB is null and NPT4_PRIV is null)
then  AVG(NPT4_OTHER) end,0) NPT4
,ifnull(case when(NPT41_OTHER is null and NPT41_PRIV is null and NPT41_PROG
is null) then  AVG(NPT41_PUB)
      when(NPT41_OTHER is null and NPT41_PUB is null and NPT41_PROG is null)
then  AVG(NPT41_PRIV)
      when(NPT41_OTHER is null and NPT41_PUB is null and NPT41_PRIV is null)
then  AVG(NPT41_PROG)
    when(NPT41_PROG is null and NPT41_PUB is null and NPT41_PRIV is null)
then  AVG(NPT41_OTHER) end,0) NPT41
,ifnull(case when(NPT42_OTHER is null and NPT42_PRIV is null and NPT42_PROG
is null) then  AVG(NPT42_PUB)
      when(NPT42_OTHER is null and NPT42_PUB is null and NPT42_PROG is null)
then  AVG(NPT42_PRIV)
      when(NPT42_OTHER is null and NPT42_PUB is null and NPT42_PRIV is null)
then  AVG(NPT42_PROG)
    when(NPT42_PROG is null and NPT42_PUB is null and NPT42_PRIV is null)
then  AVG(NPT42_OTHER) end,0) NPT42
,ifnull(case when(NPT43_OTHER is null and NPT43_PRIV is null and NPT43_PROG
is null) then  AVG(NPT43_PUB)
      when(NPT43_OTHER is null and NPT43_PUB is null and NPT43_PROG is null)
then  AVG(NPT43_PRIV)
      when(NPT43_OTHER is null and NPT43_PUB is null and NPT43_PRIV is null)
then  AVG(NPT43_PROG)
      when(NPT43_PROG is null and NPT43_PUB is null and NPT43_PRIV is null)
then  AVG(NPT43_OTHER) end,0) NPT43
,ifnull(case when(NPT44_OTHER is null and NPT44_PRIV is null and NPT44_PROG
is null) then  AVG(NPT44_PUB)
      when(NPT44_OTHER is null and NPT44_PUB is null and NPT44_PROG is null)
then  AVG(NPT44_PRIV)
      when(NPT44_OTHER is null and NPT44_PUB is null and NPT44_PRIV is null)
then  AVG(NPT44_PROG)
      when(NPT44_PROG is null and NPT44_PUB is null and NPT44_PRIV is null)
then  AVG(NPT44_OTHER) end,0) NPT44
,ifnull(case when(NPT45_OTHER is null and NPT45_PRIV is null and NPT45_PROG
is null) then  AVG(NPT45_PUB)
      when(NPT45_OTHER is null and NPT45_PUB is null and NPT45_PROG is null)
then  AVG(NPT45_PRIV)
```

```
      when(NPT45_OTHER is null and NPT45_PUB is null and NPT45_PRIV is null)
then  AVG(NPT45_PROG)
      when(NPT45_PROG is null and NPT45_PUB is null and NPT45_PRIV is null)
then  AVG(NPT45_OTHER) end,0) NPT45
,ifnull(case when(NPT4_048_OTHER is null and NPT4_048_PRIV is null and
NPT4_048_PROG is null) then  AVG(NPT4_048_PUB)
      when(NPT4_048_OTHER is null and NPT4_048_PUB is null and NPT4_048_PROG
is null) then  AVG(NPT4_048_PRIV)
      when(NPT4_048_OTHER is null and NPT4_048_PUB is null and NPT4_048_PRIV
is null) then  AVG(NPT4_048_PROG)
      when(NPT4_048_PROG is null and NPT4_048_PUB is null and NPT4_048_PRIV
is null) then  AVG(NPT4_048_OTHER) end,0) NPT4_048
,ifnull(case when(NPT4_3075_OTHER is null and NPT4_3075_PRIV is null and
NPT4_3075_PROG is null) then  AVG(NPT4_3075_PUB)
      when(NPT4_3075_OTHER is null and NPT4_3075_PUB is null and
NPT4_3075_PROG is null) then  AVG(NPT4_3075_PRIV)
      when(NPT4_3075_OTHER is null and NPT4_3075_PUB is null and
NPT4_3075_PRIV is null) then  AVG(NPT4_3075_PROG)
      when(NPT4_3075_PROG is null and NPT4_3075_PUB is null and
NPT4_3075_PRIV is null) then  AVG(NPT4_3075_OTHER) end,0) NPT4_3075
,ifnull(case when(NPT4_75UP_OTHER is null and NPT4_75UP_PRIV is null and
NPT4_75UP_PROG is null) then  AVG(NPT4_75UP_PUB)
      when(NPT4_75UP_OTHER is null and NPT4_75UP_PUB is null and
NPT4_75UP_PROG is null) then  AVG(NPT4_75UP_PRIV)
      when(NPT4_75UP_OTHER is null and NPT4_75UP_PUB is null and
NPT4_75UP_PRIV is null) then  AVG(NPT4_75UP_PROG)
      when(NPT4_75UP_PROG is null and NPT4_75UP_PUB is null and
NPT4_75UP_PRIV is null) then  AVG(NPT4_75UP_OTHER) end,0) NPT4_75UP

                       FROM Scorecard

                          WHERE
                          SCH_DEG=3 and

                          LOWER(INSTNM) IN

                          (select LOWER(INSTNM)

                          from Scorecard

                          group by LOWER(INSTNM),ZIP

                          HAVING (SUM(CASE WHEN SAT_AVG_ALL IS NULL THEN 0
ELSE 1 END) > 0))

                          GROUP BY LOWER(INSTNM) || '-' || ZIP")
```

null values replaced by mean of the field.

```
for(i in 1:ncol(trainSummary)){
  trainSummary[,i][is.na(trainSummary[,i])] <- mean(trainSummary[,i], na.rm =
TRUE)
}

## Warning in mean.default(trainSummary[, i], na.rm = TRUE): argument is not
## numeric or logical: returning NA
```

Cosine similarity function

```
getCosine<-function(x,y)
  {
  this.cosine=sum(x*y)/(sqrt(sum(x*x))*sqrt(sum(y*y)))
  return (this.cosine)
  }
```

Initial testing of the function

```
institution= trainSummary[,1]
trainSum=trainSummary[-c(1)]
insti_916=trainSum[916,]
insti_2=trainSum[2,]
insti_18=trainSum[18,]
test_inst=trainSum[4977,]
getCosine(insti_2,insti_2)

## [1] 1

getCosine(insti_2,insti_18)

## [1] 0.8405831
```

Testing for a average student of 'Yale university'(tested both for the best and the worst match)

```
sim_vec=c()
for (i in 1:nrow(trainSum)){
     sim_vec[i]=getCosine(test_inst,trainSum[i,])
}
res_data=NULL
res_data=data.frame(cbind(institution,sim_vec))

head(res_data[order(-sim_vec),],20)

##                                              institution           sim_vec
## 4977                                yale university-6520                 1
## 4364               university of pennsylvania-19104-6303 0.995879399628014
## 3553                           stanford university-94305 0.995489372553037
## 2978                     princeton university-08544-0070 0.993592279670898
## 4762                        wellesley college-02481-8203 0.992624185239721
## 1358                         georgetown university-20057 0.992415618736709
## 3069                          rice university-77005-1827 0.991874739686554
## 509             california institute of technology-91125 0.991834164229085
## 2262                            middlebury college-5753 0.990957747670238
## 2923                          pitzer college-91711-6101 0.990217217449577
## 1323 franklin w. olin college of engineering-02492-1200 0.989575038147774
## 831   columbia university in the city of new york-10027 0.989559579170433
## 4065                        university of chicago-60637 0.989316107936368
## 994                        dartmouth college-03755-3529 0.989306723419619
## 482                             bryn mawr college-19010 0.988649912532125
## 1488                            harvard university-2138 0.988269859481424
## 747                            colby college-04901-8840 0.987858175897524
## 1672                johns hopkins university-21218-2688  0.98772484745206
## 1069                              duke university-27708  0.98769057936901
## 3957                        tufts university-02155-5555  0.98761509724511
```

```
head(res_data[order(sim_vec),],10)
```

```
##                                              institution           sim_vec
## 322                             berea college-40404-2182 0.416920564726932
## 3152               sacred heart major seminary-482061799 0.447975307322105
## 3182                           saint johns seminary-2135 0.447975307361975
## 4749                                webb institute-11542 0.447975307389241
## 3529                     st johns seminary college-93012 0.447984340032563
## 3185                     saint johns seminary-93012-2598 0.447984340170487
## 3414 southeastern baptist theological seminary-275881889 0.454892605485307
## 3149                   sacred heart major seminary-48206 0.456060454478632
## 4752                            webb institute-115421398 0.459256396705151
## 4383         university of puerto rico-aguadilla-6040160   0.6545603409795
```

Conclusion-

1) For this particular testing we can found some similar university as 'Yale' but not all.
2) Other relevant fields need to consider.
3) Null value replacement with mean is very simplistic assumption. Need better approach.
4) Need to try other similarity methods, such as Pearson correlation, Euclidean,Bayesian etc.