Project Phase 5

Foo Zhi Yuan, Chew Ming Chan, Yi Ding

12/15/2016

1. How did you combine the two tables A and B to obtain E? Did you add any other table? When you did the combination, did you run into any issues? Discuss the combination process in detail, e.g., when you merge tuples, what are the merging functions (such as to merge two age values, always select the age value from the tuple from Table A, unless this value is missing in which case we select the value from the tuple in Table B).

Table E consists of the matched table we obtained from project stage 5 (call this table M), the unmatched tuples in table A (IMDB) and unmatched tuples in table B (rottentomatoes). We create table E by following the sequences below:

i.

Recall that in stage 5, we apply our matcher that is being trained on all the golden data on all of the tuples in tableC (the table that we generated through blocking), obtaining a matched table, in which we will refer to as table M. Table M has the attributes 'id', 'ltable_id', 'rtable_id', 'ltable_title', 'ltable_category', 'ltable_duration', 'ltable_rating', 'ltable_ratingCount', 'ltable_director', 'ltable_year', 'rtable_title', 'rtable_category', 'rtable_duration', 'rtable_rating', 'rtable_ratingCount', 'rtable_director' and 'rtable_year'. However, in our original table A (IMDB) and table B (rottentomatoes), we actually have a lot more attributes than these. During our blocking stage, we trim some of the attributes in order to have a uniform schema for table C and to reduce the time complexity when performing debug operation. Thus, now, we need to **add back the attributes that we trimmed off during blocking and matching stage back into table M** For IMDB, we add back the attributes "writers","date in theaters", "language", "country", "filming location", "actors", "movie description" and "story". For rottentomatoes, we add back the attributes "movie rating", "writers", "date in theaters", "date on DVD", "box office earnings", "studios", "actors" and "movie description". Then, we populate all of these attributes with their respective data.

ii. Merge the ltable and rtable in table M together.

For attributes that is available in ltable but not available in rtable or vice versa, this is a fairly easy thing to do. For example, IMDB has the attribute "story" while rottentomatoes does not. So, table M will have the attribute "story" with the data being selected from IMDB. However, for attributes that exist both in ltable and rtable, we will need to make the decision on which data to choose, IMDB or rottentomatoes? Or do we take the mean? Or do we take the string with longest length? Or do we check if the value in rtable exists in ltable, and if not, we add into it? The attributes that overlap both in ltable and rtable are: title, category, duration, rating, ratingCount, directors, year, writers, dateInTheaters, actors and movie description.

a.

For the attribute "title", we will compare the length of the strings (movie title) from ltable and rtable and pick the longest string. The reasoning behind this is we believe that the longer string typically should be a more complete movie title.

b.

For the attributes "category", "directors", "writers" and "actors", given a string X1 that contains all of the directors (but without the loss of generality) from ltable and another string X2 that contains all of the directors from rtable, we will tokenize the string X1 and X2 by comma. For each j in X2, we will check if it matches with any i in X1. If it doesn't, we will add j into X1.

c.

For the attributes "duration" and "rating", we will add the value from ltable and rtable, then divide it by 2.

d.

For the attribute "ratingCount", we added up the value from ltable and rtable.

e.

For the attribute "year", because during initially during project stage 1, we didn't extract the "year" information out of the HTML that we crawled for IMDB, and we obtained it by extracting it from the "dateInTheaters" attribute, thus, we decided that the "year" attribute for IMDB is less accurate. Therefore, given a year value from ltable (IMDB) and rtable(rotten tomatoes), as long as the year value from rtable is not empty, we will always return the year value from rtable. If rtable has empty year value, then only we will return year value from ltable.

f.

For the attribute "dateInTheaters", we wrote a method that can check if the value from ltable and rtable conform to the format "MM/DD/YEAR", because some of the date in ltable is of the format "MM/DD/YEAR" and some of the date in rtable is of the format "YEAR" or even "MONTH, YEAR" and vice versa. We want the value that is the most complete. Thus, we wrote a method, when given date from ltable and rtable, will always pick the one that conform to the format "MM/DD/YEAR". If they both conform, we will pick the date from IMDB, because IMDB is a more prestigious, trusted and famous movie review website in compared to rottentomatoes.

```
def chooseDate(x,y):
        result1=None  #IMDB
        result2=None  #rottentomatoes

        #If x is none, y is not none
        if x is None or x=='N/A' or x==":
                if y is not None and y!='N/A' and y!=":
                        return y
                else:
                        return x
        #If y is None, x is not none
        if y is None or y=='N/A' or y==":
                if x is not None and x!='N/A' and x!=":
                        return x
                else:
                        return x
        #if x and y is not None
        try:
                datetime.datetime.strptime(x,"%m/%d/%Y")
                result1=True
        except ValueError:
                result1=False

        try:
                datetime.datetime.strptime(y,"%m/%d/%Y")
                result2=True
        except ValueError:
                result2=False

        if(result1):   #give priority to IMDB
                return x
        elif(result2):
                return y
        else:
                temp=title(x,y)
                return temp
```

g.

Lastly, for the attribute “description”, we will always pick the description from ltable(IMDB) if it’s available. This is because IMDB has the attributes “description” and “story”, while rottentomatoes only has the attribute “story”. “Description” and “story” are both very similar thing, but also quite di˙erent in that “story” is more detailed. We don’t want to pick the “description” from rottentomatoes into our table M only to find out that it is the same or almost the same as the “story” attributes from IMDB (which rottentomatoes doesn’t have).

iii.For tuples that is not matched in tableA and tuples that is not matched in tableB, we append them at the ending part of table M.

Due to the fact that the schema of tableM is the union of tableA and tableB, it is expected that for tuples that are not matched in tableA and tableB, they will have a lot of missing values in table M.

2.Statistics on Table E: specifically, what is the schema of Table E, how many tuples are in Table E? Give at least four sample tuples from Table E.

i. Schema of table E, how many tuples are in Table E?

There are 21 attributes out there:

ItableID(id,indicating where this tuple is on tableA)

rtableID(id,indicating where this tuple is on tableB)

title(textual)

category(catergorical)

duration(numeric)

rating(numeric)

ratingCount(numeric)

directors(textual)

year(categorical/numeric)

movieRating(categorical)

writers(character)

dateInTheaters

language(categorical)

country(categorical)

filmingLocation(textual)

actors(textual)

dateOnDVD

boxOfficeEarnings(numeric)

studios(categorical)

movieDescription(textual)

story(textual)

And there are 36927 tuples in tableE

ii. Sample tuples:

tuple 1:4:

```
##   ltableID rtableID                        title
## 1        0     1225       Ouija: Origin of Evil
## 2        1        9              Doctor Strange
## 3        2      114 Keeping Up with the Joneses
## 4        3      176                Ghostbusters
##                                       category duration rating
## 1                                Horror,Mystery     99.0 69.00
## 2        Action,Adventure,Science Fiction,Fantasy    122.5 85.50
## 3                                 Comedy,Action    103.0 58.00
## 4 Action,Adventure,Comedy,Science Fiction,Fantasy    110.5 58.00
##   ratingCount        directors year
## 1       11085    Mike Flanagan 2016
## 2       81487 Scott Derrickson 2016
## 3        9946     Greg Mottola 2016
## 4      216657        Paul Feig 2016
##                                                                        movieRating
## 1                      PG-13 (for disturbing images, terror and thematic elements)
## 2 PG-13 (for sci-fi violence and action throughout, and an intense crash sequence)
## 3            PG-13 (for sexual content, action/violence and brief strong language)
## 4                              PG-13 (for supernatural action and some crude humor)
##                                                                                writers
## 1                                                                Mike Flanagan,Jeff Howard
## 2 Thomas Dean Donnelly,Joshua Oppenheimer,Jon Spaihts,Scott Derrickson,C. Robert Cargill
## 3                                                            Mike LeSieur,Michael LeSieur
## 4                                                                 Paul Feig,Katie Dippold
##   dateInTheaters language       country              filmingLocation
## 1     10/21/2016  English           USA Los Angeles, California, USA
## 2     11/04/2016  English           USA New York City, New York, USA
## 3     10/21/2016  English           USA        Atlanta, Georgia, USA
## 4     07/15/2016  English USA,Australia   Boston, Massachusetts, USA
##
## 1
## 2
## 3
## 4 Melissa McCarthy,Kristen Wiig,Kate McKinnon,Leslie Jones,Charles Dance,Chris Hemsworth,Michael K. W
##    dateOnDVD boxOfficeEarnings             studios
## 1 01/17/2017          34904885   Universal Pictures
## 2                    205778872 Walt Disney Pictures
## 3                     14745078     20th Century Fox
## 4 10/11/2016         128344089        Sony Pictures
##
## 1 In 1965 Los Angeles, a widowed mother and her two daughters add a new stunt to bolster their seanc
## 2
## 3
## 4
##
## 1
## 2 Marvel's "Doctor Strange" follows the story of the talented neurosurgeon Doctor Stephen Strange wh
## 3
## 4
```

3.What was the data analysis task that you wanted to do? For that task, describe in detail the data analysis process that you went through.

i. Multiple Linear Regression

We want to see whether we can use other numeric variables to predict the rating score every film has. In order to do this, we split the data into two 2 sets, the training (tuple 1:30000) and test set(tuple30001:36927).

The predictors we chose are duration,ratingCount,year,boxofficeEarning. We start by fitting the model with all 4 predictors, train the model on training data set and then apply them on the test set, to compute the accuracy. Then we try to change the predictors,delete some unsignificant predictors to see which model has the smallest error on the test data set.

ii. OLAP-exploration

We apply OLAP on the rating, specifically using slice, to see what is the highest rating movie category, who is the most successful dirctor, whether movies containing sexual contents(R rating) are more popular? Which studio provides the best movies? Which language of movie has the highest rating? Then we pick the most influential attributes and run dice on them, to see which combination of the attributes gives on average the highest movie rating.

4. Give any accuracy numbers that you have obtained (such as precision and recall for your classification scheme).

i.The result of regression analysis :

```
##
## Call:
## lm(formula = training$rating ~ training$duration + training$year +
##     training$ratingCount + training$boxOfficeEarnings)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -61.850  -5.529   0.086   5.892  26.264
##
## Coefficients:
##                              Estimate Std. Error t value Pr(>|t|)
## (Intercept)                -1.092e+02  6.901e+01  -1.582   0.1139
## training$duration           1.019e-01  1.145e-02   8.896  < 2e-16 ***
## training$year               8.209e-02  3.429e-02   2.394   0.0168 *
## training$ratingCount       -7.669e-08  5.154e-08  -1.488   0.1369
## training$boxOfficeEarnings  1.762e-08  2.643e-09   6.669 3.54e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.504 on 1606 degrees of freedom
##   (16852 observations deleted due to missingness)
## Multiple R-squared:  0.1035, Adjusted R-squared:  0.1012
## F-statistic: 46.33 on 4 and 1606 DF,  p-value: < 2.2e-16
```

we can see that year and ratingcounts are not significant as predictors

```
##
## Call:
## lm(formula = training$rating ~ training$duration + training$year +
##     training$boxOfficeEarnings)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -61.914  -5.495   0.042   5.865  26.838
##
## Coefficients:
##                               Estimate Std. Error t value Pr(>|t|)
## (Intercept)                 -1.240e+02  6.831e+01  -1.816  0.06955 .
## training$duration            1.015e-01  1.145e-02   8.863  < 2e-16 ***
## training$year                8.951e-02  3.394e-02   2.637  0.00843 **
## training$boxOfficeEarnings  1.661e-08  2.555e-09   6.502 1.05e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.508 on 1607 degrees of freedom
##   (16852 observations deleted due to missingness)
## Multiple R-squared:  0.1022, Adjusted R-squared:  0.1006
## F-statistic: 60.99 on 3 and 1607 DF,  p-value: < 2.2e-16
```

after getting rid of voting counts, year become slightly more significant but still not as good as the other 2

```
##
## Call:
## lm(formula = training$rating ~ training$duration + training$boxOfficeEarnings)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -61.435  -5.471   0.017   5.877  26.302
##
## Coefficients:
##                               Estimate Std. Error t value Pr(>|t|)
## (Intercept)                 5.608e+01  1.207e+00  46.456  < 2e-16 ***
## training$duration           9.893e-02  1.143e-02   8.653  < 2e-16 ***
## training$boxOfficeEarnings 1.676e-08  2.559e-09   6.551  7.7e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.523 on 1608 degrees of freedom
##   (16852 observations deleted due to missingness)
## Multiple R-squared:  0.09834,    Adjusted R-squared:  0.09722
## F-statistic: 87.69 on 2 and 1608 DF,  p-value: < 2.2e-16
```

The mean squared error(a measure of model accuracy) for the 3 models are showd below:

```
## [1] "model1"
```

```
## [1] 158.9805
```

```
## [1] "model2"
```

```
## [1] 158.7185
```

```
## [1] "model3"
```

```
## [1] 158.5225
```

We can see that the model with only duration and boxofficeEarning as predictor is the most accurate(has the smallest mean squared error)

ii. The result of the OLAP:

1:slice by category,compute the mean rating,showing the top 10:

```
##                              Musical & Performing Arts,Action & Adventure
##                                                                     97.00
## Action & Adventure,Animation,Art House & International,Special Interest
##                                                                     96.00
##                                Documentary,Kids & Family,Special Interest
##                                                                     94.00
##                                                          Drama,Short,War
##                                                                     93.00
##                                                 Mystery & Suspense,Short
##                                                                     91.00
##                 Art House & International,Drama,Kids & Family,Television
##                                                                     90.00
##                                      Documentary,Special Interest,Romance
##                                                                     90.00
##                                   Drama,Western,Science Fiction & Fantasy
##                                                                     89.75
##                                        Romance,History,Action & Adventure
##                                                                     89.25
##                                  Action,Adventure,Drama,Biography,Romance
##                                                                     89.00
```

2: slice by director,compute the mean rating, showing the top 10:

```
##                  Davi Russo              Patrick Suite
##                         100                        100
##                Sylvie Rokab                Tim Carroll
##                         100                        100
##                Hugh Martin Adam Brodie,Dave Derewlany
##                          99                         98
##   Ari J. Issler,Ben Snyder          Bank Tangjaitrong
##                          98                         98
##                 Bob Forward           Hettie Macdonald
##                          98                         98
```

3: slice by movieRate(restriction),compute the mean rating,showing the top 10:

```
##                       R (for violence, pervasive language, some sexual content and drug use)
##                                                                                          96
##                          NR (for language, a disturbing image, brief sexuality and drug use)
##                                                                                          94
##              PG-13 (for thematic elements involving bullying, and for brief strong language)
##                                                                                          94
```

```
##             PG (for fantasy action violence, language, some thematic material and smoking.)
##                                                                                          94
## PG (for thematic elements involving bullying and adolescent issues, and for brief language)
##                                                                                          94
##                                        PG (for thematic elements, some language and smoking)
##                                                                                          94
##                                      PG-13 (for brief suggestive humor and drug references)
##                                                                                          92
##                                      PG-13 (for some drug material, sexuality and language)
##                                                                                          90
##                                        R (for sexual material, language and brief drug use)
##                                                                                          90
##                                   PG-13 (for intense sequences of violence and some menace)
##                                                                                          89
```

4: slice by language,compute the mean rating,showing the top 10:

```
##              English,Mandarin,Japanese
##                                   96.0
## English,Dutch,French,German,Lithuanian
##                                   95.0
##                  German,German,English
##                                   95.0
##                 English,Swedish,Danish
##                                   94.5
##         English,Greek,Mandarin,Spanish
##                                   94.0
##               English,Italian,Romanian
##                                   94.0
##              English,Italian,Ukrainian
##                                   94.0
##            English,Azerbaijani,Russian
##                                   93.0
## English,Navajo,Japanese,Italian,German
##                                   93.0
##                 English,Persian,French
##                                   93.0
```

5: slice by studio,compute the mean rating,showing the top 10:

```
## In The Light Entertainment           Beech Hill Films
##                        100                         98
##             Creative Breed            Ironbound Films
##                         96                         96
##        Attention Era Media           BUCK Productions
##                         94                         94
##         Hammer Productions           Iconoclast Films
##                         94                         94
##           Musa Productions        Body Image Movement
##                         94                         92
```

6: dice by category, movieRate :the best is:

Musical & Performing Arts,Action & Adventure, with PG (for sequences of sci-fi action violence and peril, thematic elements, and language), having mean rating of 97

6: dice by category, language: the best is:

"Animation,Classics,Comedy,Horror,Kids & Family", with "English,Icelandic,Russian,French",having the mean rating of 98.

5. What did you learn/conclude from your data analysis? Were there any problems with the analysis process and with the data?

i. Conclusion:

We found that the longer the duration of the film is, the more likey it will be rated with high score. And the more earnings the film got, the more likely it will be rated with high scores, which is a little bit obvious.

And we found the categories that are rated the highest are: Musical & Performing Arts,Action & Adventure,Action & Adventure,Animation,Art House & International,Special Interest, and Documentary,Kids & Family,Special Interest.

Also, people tend to like :movies that are labeled R (for violence, pervasive language, some sexual content and drug use) the most. Well we love strong stimulations, don't we?

And the best films are in English,Mandarin,Japanese. (Well....the countries that have the largest GDP?)

ii. Problems:

We learnt that regression analysis is having somewhat high demand of our data. First of all, it requires our data to be as continous as possible. Our year data is like 2012,2013,2014, which is not that continous, resulting in the result of years as predictor not very satisfying. Also, it is really hard to include all possible power of predictors. For example, we have four predictors, if we include (1.) not include (2.)itself (3.its square) (4. its cubic) for each of the four predictor, then we have 4^4=256 models to compute, which is really an impossible task if computing each model and predict it will already consume a certain amount of time.

Also, our data is taking certain combination of word (e.g. English,Chinsese,Japanese)as a possible value for 'language' attribute, instead of listing out all the possible single languages and indicate which languages each tuple has. This might require further attribute breaking down but will certainly help more. Mainwhile, it is a bit hard to save data that are easy do this kind of analysis and operation (e.g. OLAP) in the form of CSV, as it is hard for csv to handle 3 or more dimensional data.

6 If you have more time, what would you propose you can do next?

If we have more time, we will probably further break up and rearrange some of our attributes that is in the form of a combination of some units, so that it will be a lot easier for us to do OLAP-styled analysis.

Also, we will need more time to handle the massive textual data for every film that is describing its story or so on.

Making a list of all frequently casted actors or actress will also be interesting to do, and again it needs a deeper data manipulation on the attribute 'actors'.

Another thing that is on our mind is to take full advantage of the attribute 'film location'. If possible, it will be great fun to plot those data into a map, to give others a more intuitive and visualized feeling of the distribution of popular films spots all over the world.