Final Project - Yelp Recommender System

Juliann McEachern, Rajwant Mishra,Christina Valore

July 16, 2019

Overview

Add overview here.

```
## Warning: package 'dplyr' was built under R version 3.5.3

##
## 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

## Warning: package 'tidyr' was built under R version 3.5.3

## Warning: package 'jsonlite' was built under R version 3.5.3

## Warning: package 'kableExtra' was built under R version 3.5.3

##
## Attaching package: 'kableExtra'

## The following object is masked from 'package:dplyr':
##
##     group_rows

## Warning: package 'default' was built under R version 3.5.2

## Warning: package 'ggplot2' was built under R version 3.5.3

## Warning: package 'recommenderlab' was built under R version 3.5.3

## Loading required package: Matrix

## Warning: package 'Matrix' was built under R version 3.5.3

##
## Attaching package: 'Matrix'
```

```
## The following object is masked from 'package:tidyr':
##
##     expand

## Loading required package: arules

## Warning: package 'arules' was built under R version 3.5.3

##
## Attaching package: 'arules'

## The following object is masked from 'package:dplyr':
##
##     recode

## The following objects are masked from 'package:base':
##
##     abbreviate, write

## Loading required package: proxy

## Warning: package 'proxy' was built under R version 3.5.2

##
## Attaching package: 'proxy'

## The following object is masked from 'package:Matrix':
##
##     as.matrix

## The following objects are masked from 'package:stats':
##
##     as.dist, dist

## The following object is masked from 'package:base':
##
##     as.matrix

## Loading required package: registry

## Warning: package 'registry' was built under R version 3.5.2

## Warning: package 'Metrics' was built under R version 3.5.3

## Warning: package 'lsa' was built under R version 3.5.3

## Loading required package: SnowballC

## Warning: package 'SnowballC' was built under R version 3.5.2

## Warning: package 'diveRsity' was built under R version 3.5.3
```

Data Aquisition

@ Raj - Add textual explanation of process. What data is available? Explain variables we have to work with.

Data Transformations

@ Raj - Explain process. What was included/excluded? Why? Ie:

We changed the business and user columns name for clarity to better identify our variables.

We then joined our dataframes using the `Business_ID` and `User_ID` as unique keys.

```
#-------------------------------------------------------------------------------------------
# Building Main data set by Joining all the above data sets on key of Business_ID and User_ID
#-------------------------------------------------------------------------------------------

#Group all reviews that have a business with a star rating
review_bus  <- inner_join(df_review,df_business,by=("business_id"))

#Group all reviews that have and User's data
review_user <- inner_join(df_review,df_user,by=("user_id"))

# joining all the "Review and Users" information with  the "Review and Business"  Dataset.
# Moving Key in the Begning
df_main <- inner_join(review_bus,review_user[,-c(2,3,4,5,6,8,9,10)],
                              by =c("business_id","user_id")) %>% .[,c(1,2,7,3,4,5,6,8:26)]
# Appending the Checkins information for the Business
df_main_chk <- inner_join(df_main,df_checkin[,-c(1)],by=("business_id"))
```

Data Introduction

Business

This dataset list business information with average business rating, along with category of the Business.

Business Train Data

```
head(df_business)
```

```
##      bus_city            business_id
## 1      Peoria rncjoVoEFUJGCUoC1JgnUA
## 2     Phoenix 0FNFSzCFP_rGUoJx8W7tJg
## 3     Phoenix 3f_lyB6vFK48ukH6ScvLHg
## 4     Phoenix usAsSV36QmUej8--yvN-dg
## 5 Glendale Az PzOqRohWw7F7YEPBz6AubA
## 6    Glendale gtQzAiy7D-dPU8WzT3jX3Q
##                                                        bus_categories
## 1 Accountants, Professional Services, Tax Services, Financial Services
## 2                                      Sporting Goods, Bikes, Shopping
## 3
## 4                                                         Food, Grocery
## 5                                     Food, Bagels, Delis, Restaurants
```

```
## 6                                  Women's Clothing, Fashion, Shopping
##                       bus_name bus_open bus_review_count bus_state
## 1    Peoria Income Tax Service     TRUE                3        AZ
## 2                  Bike Doctor     TRUE                5        AZ
## 3 Valley Permaculture Alliance     TRUE                4        AZ
## 4                    Food City     TRUE                5        AZ
## 5            Hot Bagels & Deli     TRUE               14        AZ
## 6      Barney's New York Co-op     TRUE                6        CA
##   bus_type bus_stars
## 1 business       5.0
## 2 business       5.0
## 3 business       5.0
## 4 business       3.5
## 5 business       3.5
## 6 business       4.5
##                                                    bus_categories_all
## 1 Accountants, Professional Services, Tax Services, Financial Services
## 2                                     Sporting Goods, Bikes, Shopping
## 3
## 4                                                         Food, Grocery
## 5                                    Food, Bagels, Delis, Restaurants
## 6                                  Women's Clothing, Fashion, Shopping
```

Review

Review is main dataset here, it holds link to Users and Business and the corresponding Text Reviews are listed along with ratings.

Review Train Data

```
head(df_review)
```

```
##                  user_id              review_id rev_stars   rev_date
## 1 rLtl8ZkDX5vH5nAx9C3q5Q fWKvX83p0-ka4JS3dc6E5A         5 2011-01-26
## 2 0a2KyEL0d3Yb1V6aivbIuQ IjZ33sJrzXqU-0X6U8NwyA         5 2011-07-27
## 3 0hT2KtfLiobPvh6cDC8JQg IESLBzqUCLdSzSqm0eCSxQ         4 2012-06-14
## 4 uZetl9T0NcROGOyFfughhg G-WvGaISbqqaMHlNnByodA         5 2010-05-27
## 5 vYmM4KTsC8ZfQBg-j5MWkw 1uJFq2r5QfJG_6ExMRCaGw         5 2012-01-05
## 6 sqYN3lNgvPbPCTRsMFu27g m2CKSsepBCoRYWxiRUsxAg         4 2007-12-13
##
## 1
## 2
## 3
## 4
## 5
## 6 Quiessence is, simply put, beautiful.  Full windows and earthy wooden walls give a feeling of warm
##   rev_type            business_id rev_funny rev_useful rev_cool
## 1   review 9yKzy9PApeiPPOUJEtnvkg         0          5        2
## 2   review ZRJwVLyzEJq1VAihDhYiow         0          0        0
## 3   review 6oRAC4uyJCsJl1X0WZpVSA         0          1        0
## 4   review _1QQZuf4zZOyFCvXc0o6Vg         0          2        1
## 5   review 6ozycU1RpktNG2-1BroVtw         0          0        0
## 6   review -yxfBYGB6SEqszmxJxd97A         1          3        4
```

User

Users Holds all the information about user, average rating of the Users given so far, count of the review and different count of Votes on “Cool”, “Useful” and “funny”.

User Train Data

```
head(df_user)
```

```
##                  user_id  usr_name usr_review_count usr_type
## 1 CR2y7yEm4X035ZMzrTtN9Q       Jim                6     user
## 2 _9GXoHhdxc30ujPaQwh6Ew     Kelle                2     user
## 3 8mM-nqxjg6pT04kwcjMbsw Stephanie                2     user
## 4 Ch6CdTR2IVaVANr-RglMOg         T                2     user
## 5 NZrLmHRyiHmyT1JrfzkCOA      Beth                1     user
## 6 mWx5Sxt_dx-sYBZg6RgJHQ       Amy               19     user
##   usr_average_stars usr_funny usr_useful usr_cool
## 1              5.00         0          7        0
## 2              1.00         0          1        0
## 3              5.00         0          1        0
## 4              5.00         0          2        0
## 5              1.00         0          0        0
## 6              3.79        30         45       36
```

Checkin

Checkin information is round the clock and 7 days data of the Customer’s checkin. Column name start with hour-Day here hours [00 to 23]- [0 to 6]

Checkin Train Data

```
head(df_checkin)
```

```
##      type            business_id 11-3 8-5 15-0 15-3 15-5 14-4 14-5 14-6
## 1 checkin KO9CpaSPOoqm0iCWm5scmg   17   1    2    2    2    1    3    6
## 2 checkin oRqBAYtcBYZHXA7G8FlPaA   NA  NA   NA   NA   NA   NA   NA    1
## 3 checkin 6cy2C9aBXUwkrh4bY1DApw    1   1   NA    1   NA   NA   NA    1
## 4 checkin D0IB17N66FiyYDCzTlAI4A   NA  NA   NA   NA   NA   NA   NA   NA
## 5 checkin HLQGo3EaYVvAv22bONGkIw   NA  NA   NA   NA   NA   NA    1   NA
## 6 checkin J6OojF0R_1OuwNlrZI-ynQ   NA  NA   NA   NA    1   NA   NA   NA
##   14-0 14-1 14-3 0-5 1-6 11-5 11-4 13-1 11-6 11-1 13-6 13-5 11-2 12-6 12-4
## 1    2    2    2   1   1    3   11    1    6   18    5    4    9    5    8
## 2   NA   NA   NA   1   2   NA   NA   NA   NA   NA    1   NA    1    1   NA
## 3   NA   NA   NA  NA  NA    4   NA   NA    5   NA   NA   NA   NA   NA   NA
## 4   NA   NA   NA  NA  NA    2   NA   NA   NA    2   NA    1   NA    1   NA
## 5   NA   NA   NA  NA  NA   NA   NA   NA   NA   NA   NA   NA   NA   NA   NA
## 6   NA   NA   NA  NA  NA   NA   NA   NA   NA   NA   NA   NA   NA   NA   NA
##   12-5 12-2 12-3 12-0 12-1 13-3 9-5 9-4 13-2 20-1 9-6 16-3 16-1 16-5 10-0
## 1    5   12   19   20   14    1   2   1    6    1   4    1    1    1    3
## 2   NA   NA   NA   NA   NA   NA  NA  NA   NA   NA  NA   NA   NA   NA   NA
## 3   NA   NA   NA   NA    1   NA   1  NA   NA   NA  NA    1   NA   NA    1
## 4    1   NA    1   NA   NA   NA  NA  NA    1   NA  NA   NA   NA   NA    1
## 5    2   NA   NA   NA   NA   NA  NA  NA   NA   NA  NA   NA   NA   NA   NA
```

```
## 6   NA   NA   NA   NA   NA   NA  NA  NA   NA   NA  NA   NA   NA   NA   NA
##   10-1 10-2 10-3 10-4 10-5 10-6 11-0 2-6 2-5 3-6 3-5 19-4 19-2 18-0 23-0
## 1    4    4    4    1    2    2    3  NA  NA  NA  NA   NA   NA   NA   NA
## 2   NA   NA   NA   NA   NA    1   NA   2   3   1   1    1    1    1    1
## 3    1   NA    1    1    1    2   NA  NA  NA  NA  NA   NA    1   NA   NA
## 4    1   NA   NA   NA    2   NA    1  NA  NA  NA  NA   NA    1   NA   NA
## 5   NA   NA   NA   NA   NA   NA   NA  NA  NA  NA  NA   NA   NA   NA   NA
## 6   NA   NA   NA   NA   NA   NA   NA  NA  NA  NA  NA   NA   NA   NA   NA
##   17-4 18-5 17-1 0-0 20-3 13-4 7-4 23-4 14-2 0-2 19-1 19-0 19-3 6-4 6-2
## 1   NA   NA   NA  NA   NA   NA  NA   NA   NA  NA   NA   NA   NA  NA  NA
## 2    1    4    1   3    1   NA  NA   NA   NA  NA   NA   NA   NA  NA  NA
## 3    1    1   NA  NA   NA    1   1    1    1   1    1    1    1   1   1
## 4   NA   NA    1  NA   NA    2  NA   NA   NA  NA   NA   NA   NA  NA  NA
## 5   NA   NA   NA  NA   NA    1  NA   NA   NA  NA   NA   NA   NA  NA  NA
## 6   NA    1   NA  NA   NA   NA  NA   NA   NA  NA   NA   NA    1  NA  NA
##   6-3 9-2 20-5 7-5 20-6 21-6 21-0 8-2 17-6 15-1 20-0 16-2 15-4 18-4 8-4
## 1  NA  NA   NA  NA   NA   NA   NA  NA   NA   NA   NA   NA   NA   NA  NA
## 2  NA  NA   NA  NA   NA   NA   NA  NA   NA   NA   NA   NA   NA   NA  NA
## 3   1   1    1   2    1    1    1   1   NA   NA   NA   NA   NA   NA  NA
## 4  NA  NA   NA  NA   NA   NA   NA  NA    1    1    1   NA   NA   NA  NA
## 5  NA  NA   NA  NA   NA   NA   NA  NA   NA   NA   NA    1    1   NA  NA
## 6  NA  NA   NA  NA   NA   NA   NA  NA   NA   NA   NA   NA   NA    1  NA
##   6-6 18-1 5-4 7-6 17-5 8-6 16-0 21-1 8-0 17-2 15-2 15-6 18-2 18-6 17-3
## 1  NA   NA  NA  NA   NA  NA   NA   NA  NA   NA   NA   NA   NA   NA   NA
## 2  NA   NA  NA  NA   NA  NA   NA   NA  NA   NA   NA   NA   NA   NA   NA
## 3  NA   NA  NA  NA   NA  NA   NA   NA  NA   NA   NA   NA   NA   NA   NA
## 4  NA   NA  NA  NA   NA  NA   NA   NA  NA   NA   NA   NA   NA   NA   NA
## 5  NA   NA  NA  NA   NA  NA   NA   NA  NA   NA   NA   NA   NA   NA   NA
## 6  NA   NA  NA  NA   NA  NA   NA   NA  NA   NA   NA   NA   NA   NA   NA
##   19-6 19-5 9-1 1-1 18-3 16-6 7-2 17-0 21-2 13-0 23-2 22-6 8-3 21-4 23-3
## 1   NA   NA  NA  NA   NA   NA  NA   NA   NA   NA   NA   NA  NA   NA   NA
## 2   NA   NA  NA  NA   NA   NA  NA   NA   NA   NA   NA   NA  NA   NA   NA
## 3   NA   NA  NA  NA   NA   NA  NA   NA   NA   NA   NA   NA  NA   NA   NA
## 4   NA   NA  NA  NA   NA   NA  NA   NA   NA   NA   NA   NA  NA   NA   NA
## 5   NA   NA  NA  NA   NA   NA  NA   NA   NA   NA   NA   NA  NA   NA   NA
## 6   NA   NA  NA  NA   NA   NA  NA   NA   NA   NA   NA   NA  NA   NA   NA
##   20-4 16-4 21-5 3-2 22-4 2-3 2-4 20-2 21-3 7-0 4-2 5-6 5-1 6-5 6-0 6-1
## 1   NA   NA   NA  NA   NA  NA  NA   NA   NA  NA  NA  NA  NA  NA  NA  NA
## 2   NA   NA   NA  NA   NA  NA  NA   NA   NA  NA  NA  NA  NA  NA  NA  NA
## 3   NA   NA   NA  NA   NA  NA  NA   NA   NA  NA  NA  NA  NA  NA  NA  NA
## 4   NA   NA   NA  NA   NA  NA  NA   NA   NA  NA  NA  NA  NA  NA  NA  NA
## 5   NA   NA   NA  NA   NA  NA  NA   NA   NA  NA  NA  NA  NA  NA  NA  NA
## 6   NA   NA   NA  NA   NA  NA  NA   NA   NA  NA  NA  NA  NA  NA  NA  NA
##   9-0 9-3 4-4 5-3 5-2 7-3 7-1 3-4 8-1 0-3 22-5 23-1 22-0 22-3 23-6 22-2
## 1  NA  NA  NA  NA  NA  NA  NA  NA  NA  NA   NA   NA   NA   NA   NA   NA
## 2  NA  NA  NA  NA  NA  NA  NA  NA  NA  NA   NA   NA   NA   NA   NA   NA
## 3  NA  NA  NA  NA  NA  NA  NA  NA  NA  NA   NA   NA   NA   NA   NA   NA
## 4  NA  NA  NA  NA  NA  NA  NA  NA  NA  NA   NA   NA   NA   NA   NA   NA
## 5  NA  NA  NA  NA  NA  NA  NA  NA  NA  NA   NA   NA   NA   NA   NA   NA
## 6  NA  NA  NA  NA  NA  NA  NA  NA  NA  NA   NA   NA   NA   NA   NA   NA
##   0-1 1-0 22-1 23-5 0-6 1-5 5-0 0-4 3-1 1-4 1-2 4-3 4-6 5-5 4-5 4-0 4-1
## 1  NA  NA   NA   NA  NA  NA  NA  NA  NA  NA  NA  NA  NA  NA  NA  NA  NA
## 2  NA  NA   NA   NA  NA  NA  NA  NA  NA  NA  NA  NA  NA  NA  NA  NA  NA
## 3  NA  NA   NA   NA  NA  NA  NA  NA  NA  NA  NA  NA  NA  NA  NA  NA  NA
```

```
## 4  NA  NA   NA   NA  NA  NA  NA  NA  NA  NA  NA  NA  NA  NA  NA  NA  NA
## 5  NA  NA   NA   NA  NA  NA  NA  NA  NA  NA  NA  NA  NA  NA  NA  NA  NA
## 6  NA  NA   NA   NA  NA  NA  NA  NA  NA  NA  NA  NA  NA  NA  NA  NA  NA
##   1-3 3-3 2-0 2-1 2-2 3-0
## 1  NA  NA  NA  NA  NA  NA
## 2  NA  NA  NA  NA  NA  NA
## 3  NA  NA  NA  NA  NA  NA
## 4  NA  NA  NA  NA  NA  NA
## 5  NA  NA  NA  NA  NA  NA
## 6  NA  NA  NA  NA  NA  NA
```

Final Data:

- We renamed All the Columns from Business (bus_),User(usr_) and Review (rev_) dataset, so that we can identify them from the Big dataset
- Review dataset will be combined with User and Business dataset, with User_ID and Business_Id as key
- Final Dataset is then combined with Checkin information of the Businesses

```
## # A tibble: 43,873 x 2
##    user_id                    n
##    <chr>                  <int>
##  1 fczQCSmaWF78toLEmb0Zsw   588
##  2 90a6z--_CUrl84aCzZyPsg   506
##  3 4ozupHULqGyO42s3zNUzOQ   442
##  4 joIzw_aUiNvBTuGoytrH7g   392
##  5 0bNXP9quoJEgyVZu9ipGgQ   376
##  6 JffajLV-Dnn-eGYgdXDxFg   371
##  7 3gIfcQq5KxAegwCPXc83cQ   340
##  8 _PzSNcfrCjeBxSLXRoMmgQ   333
##  9 ikm0UCahtK34LbLCEw4YTw   328
## 10 lPaYMDmJbAnv_3pmZH_inw   320
## # ... with 43,863 more rows
```

Data Exploration

WIP

Data Visualizations

WIP

Data Preparation for Model

Matrix Building

We converted our raw ratings data into a user-item matrix to test and train our subsequent recommender system algorithms.

```
# create user item matrix with  only as I was getting storage error.
# https://github.com/tidyverse/tidyr/issues/426



#rownames(ui_matrix_star)<-ui_matrix_star$user_id # set row names to user_id
#ui_matrix_star<-ui_matrix_star %>% select(-user_id) %>% as.matrix()# remove user_id from columns
#umat <- as(ui_matrix_star,"realRatingMatrix") # save real ratings for algo
#real_ui_matrix_star <- as(ui_matrix_star,"realRatingMatrix") # save real ratings for algo

# preview matrices
#as.data.frame.array(ui_matrix_star) %>% head() %>% kable(caption="Preview of User-Item Matrix (User-Bu
```

Training and Test Subsets

Finally, our data was split into training and tests sets for model evaluation of both two recommender algorithms. We split our data with 10 k-folds using the `recommendaerlab` package. 80% of data was retained for training and 20% for testing purposes.

```
# evaluation method with 90% of data for train and 10% for test
set.seed(1000)

#evalu <- evaluationScheme(real_ui_matrix_star, method="split", train=0.8, given=0)

# Prep data
#ratings_train <- getData(evalu, 'train')# Training Dataset
#ratings_test_known <- getData(evalu, 'known') # Test data from evaluationScheme of type KNOWN
#ratings_test_unknown <- getData(evalu, 'unknown') # Unknow datset used for RMSE / model evaluation
```

Algorithm

Conclusion

Accuracy Metrics

References

- **Tidyr Issue**
- **Data Overview**