DSI Assignment 1

Audrey Pentz PNTAUD001

September 3, 2018

Description of the problem

The general aim of any collaborative filtering system is to make recommendations to existing or new users, based on their own history of usage of items and compared with the history of usage of items of other users that are similar to them.

The terms `users`, `items` and `similarity` differ between methods and can be applied to various fields, for example: Netflix recommends movies (`items`) to subscribers (`users`) based on their self selected favourites (new users) and viewing history (existing users) and Amazon recommends books (`items`) to account holders (`users`) based on their purchase history and other account holders that have purchased the same item as well as items that account holder has not (yet) purchased.

The specific task of this assignment is to use the modified version of the `Book-Crossings` dataset provided, to make recommendations to new or existing readers (`users`), of books (`items`) they may enjoy, based on previous books they have rated. The dataset comprises book ratings of 10,000 users that have rated 150 books on a scale from 0 to 10.

There are 3 main collaborative filtering techniques that have been explored, namely: – User Based Collaborative Filtering
– Item Based Collaborative Filtering
– Collaborative Filterin using the Matrix Factorisation technique

Approach

There are many collaborative filtering packages available on CRAN and GitHub so the objective of building this recommendation system as a package is not to try to better those but to try to automate some of the repetitive tasks and the audience in mind is my future self, needing a refresh of the methods. With that in mind, 2 of the main packages that were used to build on were the `coop` package to calculate cosine similarities fast and the `NNLM` package for matrix factorisation.

A lot of work goes into building a package so it was designed so that it can be used on any similar dataset, rather than customising it to only be appropriate for the `Book-Crossings` dataset.

Many of the tasks required by the 3 methods overlap and where there was overlap, functions were built and documented to make it easier to execute the same task multiple times (rather than copying code) eg. all 3 methods require a transformation of the data so the following 3 functions were built for this purpose: – `remove_zeros_blanks` - all 3 required a removal of implicit ratings (rating = zero) and even though this dataset does not have any NA (blank) values, this function also caters for that – `long_to_wide` - all 3 required a conversion from long format to wide format however User and Item based methods required no ratings to be filled with zeros whilst matrix factorisation required no ratings to be filled with NA's so this was catered for in this function – `convert_to_matrix` - all 3 methods have this requirement

In contrast, splitting the data into a training and test set is a task only required by matrix factorisation and is only executed once per dataset so this was not built into a function.

User and Item based methods required the calculation of similarities and a recommendation method and Matrix Factorisation required model building, model evaluation, a prediction method and a recommendation method. These were all built as functions and naming convensions were chosen in such a way that the user only needs to remember which method they are trying to execute to call up the function, namely: – `user_similarities`

- user_predict
- item_similarities
- item_predict
- mf_model_build
- mf_model_rmse
- mf_model_predict
- mf_model_recommend

Please see the documentation of each function for more information about how the functions work as well as the vignettes for an end to end example of each method.

The package is called the cf package (for collaborative filtering) which should be easy enough to remember (even for my future self). The modified version of the Book-Crossings dataset has been imported into the package so that anyone can learn from this package by applying the same techniques to the same data and perhaps extend the work done.

There are many places that building a package could go wrong so the initial approach to build everything outside of the package and then convert it into the format required to build the package, did not work well. We therefore recommend an iterative approach when building a package which starts by creating a function then going through the package build cycle (document, build, restart R) and then using the function in a vignette as it's only through trying to use the package that one realises what's missing, what's not working and where there is room for improvement. The other reason for this approach is that vignettes seem to be even more prone to errors so the sooner one tackles those, the easier it is to resolve.

Results

The detailed results are contained within 4 vignettes which were built as part of the cf package and all code is exposed in order for a user of the package to be able to execute similar tasks on their dataset. These are described below:

- EDA - this is an Exploratory Data Analysis of the modified version of the Book-Crossings dataset to showcase the kind of problems a user of the package may need to resolve with their dataset and as a reminder of what this dataset contains.
- cf_user_based - this is an illustration of how to do User based collaborative filtering.
- cf_item_based - this is an illustration of how to do Item based collaborative filtering.
- cf_matrix_factorisation - this is an illustration of how to do collaborative filtering using the Matrix Factorisation technique.

To summarise the results, we chose 5 Users that have rated 5 or more books to illustrate the results. The Users are: "23872","63360","98783","16795", "11676".

Recommendations using the User-Based method:

```
## Joining, by = "ISBN"
```

```
##       User.ID       ISBN                 Book.Title Book.Rating
## 1009    11676 0312291639 the nanny diaries: a novel           9
## 16885  251140 0312291639 the nanny diaries: a novel          10
```

```
## Joining, by = c("User.ID", "ISBN", "Book.Title", "Book.Rating")
```

```
## [[1]]
##                                                                title
## 1          the queen of the damned (vampire chronicles (paperback))
## 2  harry potter and the sorcerer's stone (harry potter (paperback))
## 3                                         the lovely bones: a novel
## 4      the hobbit : the enchanting prelude to the lord of the rings
## 5                                                       jurassic park
## 6                                                   the pelican brief
## 7                                                            the firm
## 8                                                          red dragon
## 9                                               to kill a mockingbird
## 10                                             the catcher in the rye
##        score
## 1  131.05363
## 2   93.83288
## 3   89.96614
## 4   84.06274
## 5   77.17642
## 6   75.95878
## 7   74.76121
## 8   72.57856
## 9   67.74931
## 10  65.67946
##
## [[2]]
##                                                                title
## 1          the queen of the damned (vampire chronicles (paperback))
## 2      the hobbit : the enchanting prelude to the lord of the rings
## 3                                                          red dragon
## 4  harry potter and the sorcerer's stone (harry potter (paperback))
## 5                  the witching hour (lives of the mayfair witches)
## 6                                                       jurassic park
## 7                harry potter and the order of the phoenix (book 5)
## 8                                                      a time to kill
## 9                                                            the firm
## 10                                                  the da vinci code
##        score
## 1  167.27326
## 2  102.71914
## 3   74.91599
## 4   70.72672
## 5   64.54331
## 6   62.45144
## 7   52.55600
## 8   52.14719
## 9   47.05835
## 10  46.06785
##
## [[3]]
##                                                                title
## 1      the hobbit : the enchanting prelude to the lord of the rings
## 2                  the witching hour (lives of the mayfair witches)
## 3                                                silence of the lambs
## 4  harry potter and the sorcerer's stone (harry potter (paperback))
## 5                                                   the da vinci code
## 6                harry potter and the order of the phoenix (book 5)
## 7                                              the catcher in the rye
## 8                                                       jurassic park
## 9                                           the lovely bones: a novel
## 10                                                     a time to kill
##        score
## 1  103.23589
## 2   76.05146
## 3   64.50905
## 4   63.91910
## 5   52.97571
## 6   51.19824
## 7   46.03790
## 8   46.00951
## 9   44.23505
## 10  40.72549
##
## [[4]]
##                                                                         title
## 1                            divine secrets of the ya-ya sisterhood: a novel
## 2                                                 the pilot's wife : a novel
## 3          harry potter and the sorcerer's stone (harry potter (paperback))
## 4                                                  girl with a pearl earring
## 5                                                      to kill a mockingbird
## 6  tuesdays with morrie: an old man, a young man, and life's greatest lesson
## 7                                                        angels &amp; demons
## 8                                                     snow falling on cedars
## 9                          she's come undone (oprah's book club (paperback))
## 10                       harry potter and the order of the phoenix (book 5)
##       score
## 1  185.1681
## 2  162.8732
## 3  146.5021
## 4  145.4039
## 5  137.5053
## 6  135.9401
## 7  134.4965
## 8  127.5379
## 9  124.3682
## 10 121.8260
##
## [[5]]
##                                                  title    score
## 1                               the secret life of bees 191.9730
## 2     divine secrets of the ya-ya sisterhood: a novel 189.1042
## 3                                   angels &amp; demons 159.4636
## 4                            the pilot's wife : a novel 156.1551
## 5  harry potter and the order of the phoenix (book 5) 152.7398
## 6                             girl with a pearl earring 149.8742
## 7                                        a time to kill 130.0565
## 8                                 to kill a mockingbird 130.0160
## 9              she's come undone (oprah's book club) 123.8034
## 10                                   the joy luck club 119.0809
```

Recommendations using the Item-Based method:

```
## Joining, by = "ISBN"
```

```
## [[1]]
## # A tibble: 10 x 2
##    title                                                        score
##    <chr>                                                        <dbl>
##  1 the queen of the damned (vampire chronicles (paperback))     1.03 
##  2 harry potter and the chamber of secrets (book 2)             0.701
##  3 the client                                                   0.675
##  4 the pelican brief                                            0.667
##  5 jurassic park                                                0.654
##  6 the firm                                                     0.649
##  7 red dragon                                                   0.606
##  8 the hobbit : the enchanting prelude to the lord of the rings 0.590
##  9 harry potter and the order of the phoenix (book 5)           0.560
## 10 the witching hour (lives of the mayfair witches)             0.550
## 
## [[2]]
## # A tibble: 10 x 2
##    title                                                            score
##    <chr>                                                            <dbl>
##  1 the queen of the damned (vampire chronicles (paperback))         0.911
##  2 the witching hour (lives of the mayfair witches)                 0.420
##  3 the hobbit : the enchanting prelude to the lord of the rings     0.370
##  4 red dragon                                                       0.367
##  5 jurassic park                                                    0.311
##  6 a time to kill                                                   0.283
##  7 hannibal                                                         0.240
##  8 kiss the girls                                                   0.235
##  9 the firm                                                         0.234
## 10 harry potter and the sorcerer's stone (harry potter (paperback)) 0.232
## 
## [[3]]
## # A tibble: 10 x 2
##    title                                                            score
##    <chr>                                                            <dbl>
##  1 silence of the lambs                                             0.432
##  2 the hobbit : the enchanting prelude to the lord of the rings     0.412
##  3 a time to kill                                                   0.269
##  4 the catcher in the rye                                           0.251
##  5 jurassic park                                                    0.250
##  6 the bad beginning (a series of unfortunate events, book 1)       0.248
##  7 watership down                                                   0.247
##  8 harry potter and the sorcerer's stone (harry potter (paperback)) 0.239
##  9 the handmaid's tale                                              0.231
## 10 the firm                                                         0.225
## 
## [[4]]
## # A tibble: 10 x 2
##    title                                                            score
##    <chr>                                                            <dbl>
##  1 bridget jones's diary                                             4.19
##  2 i know this much is true                                          3.96
##  3 balzac and the little chinese seamstress : a novel                3.70
##  4 to kill a mockingbird                                             3.69
##  5 wicked: the life and times of the wicked witch of the west        3.66
##  6 she's come undone (oprah's book club (paperback))                 3.65
##  7 the brethren                                                      3.65
##  8 b is for burglar (kinsey millhone mysteries (paperback))          3.58
##  9 the girls' guide to hunting and fishing                           3.56
## 10 harry potter and the sorcerer's stone (harry potter (paperback))  3.55
##
## [[5]]
## # A tibble: 10 x 2
##    title             score
##    <chr>             <dbl>
##  1 girl, interrupted  4.58
##  2 <NA>              NA
##  3 <NA>              NA
##  4 <NA>              NA
##  5 <NA>              NA
##  6 <NA>              NA
##  7 <NA>              NA
##  8 <NA>              NA
##  9 <NA>              NA
## 10 <NA>              NA
```

Recommendations using the Matrix Factorisation method:

```
## Joining, by = "ISBN"
```

```
##   User.ID                            Book.Title Book.Rating
## 1  276925                     the da vinci code           8
## 2  277042     roses are red (alex cross novels)           7
## 3  277042                      violets are blue           8
## 4  277042                           wild animus           2
## 5  277195 she's come undone (oprah's book club)          10
## 6  277378   the red tent (bestselling backlist)           7
```

```
##       User.ID                 Book.Title Book.Rating
## 1009    11676 the nanny diaries: a novel           9
## 16885  251140 the nanny diaries: a novel          10
```

```
## Joining, by = c("User.ID", "Book.Title", "Book.Rating")
```

```
##   User.ID                            Book.Title Book.Rating
## 1  276925                     the da vinci code           8
## 2  277042     roses are red (alex cross novels)           7
## 3  277042                      violets are blue           8
## 4  277042                           wild animus           2
## 5  277195 she's come undone (oprah's book club)          10
## 6  277378   the red tent (bestselling backlist)           7
```

```
## [[1]]
##                                                                title
## 1                                              to kill a mockingbird
## 2                                     seabiscuit: an american legend
## 3  harry potter and the sorcerer's stone (harry potter (paperback))
## 4      the hobbit : the enchanting prelude to the lord of the rings
## 5                                                     watership down
## 6                 harry potter and the order of the phoenix (book 5)
## 7                                            a prayer for owen meany
## 8          fast food nation: the dark side of the all-american meal
## 9                                                    fahrenheit 451
## 10                                         the pillars of the earth
##       score
## 1  9.791041
## 2  9.719797
## 3  9.581948
## 4  9.559968
## 5  9.546394
## 6  9.527658
## 7  9.465885
## 8  9.442707
## 9  9.410335
## 10 9.303176
##
## [[2]]
##                                                                title
## 1                                             to kill a mockingbird
## 2                                    seabiscuit: an american legend
## 3  harry potter and the sorcerer's stone (harry potter (paperback))
## 4      the hobbit : the enchanting prelude to the lord of the rings
## 5                                                    watership down
## 6                harry potter and the order of the phoenix (book 5)
## 7                                           a prayer for owen meany
## 8          fast food nation: the dark side of the all-american meal
## 9                                                    fahrenheit 451
## 10                                             silence of the lambs
##       score
## 1  9.001251
## 2  8.935754
## 3  8.809024
## 4  8.788817
## 5  8.776339
## 6  8.759113
## 7  8.702323
## 8  8.681015
## 9  8.651254
## 10 8.582163
##
## [[3]]
##                                                                title
## 1                                             to kill a mockingbird
## 2                                    seabiscuit: an american legend
## 3  harry potter and the sorcerer's stone (harry potter (paperback))
## 4      the hobbit : the enchanting prelude to the lord of the rings
## 5                harry potter and the order of the phoenix (book 5)
## 6                                           a prayer for owen meany
## 7          fast food nation: the dark side of the all-american meal
## 8                                              silence of the lambs
## 9                                          the pillars of the earth
## 10                 harry potter and the chamber of secrets (book 2)
##       score
## 1  8.810626
## 2  8.746515
## 3  8.622470
## 4  8.602690
## 5  8.573616
## 6  8.518028
## 7  8.497171
## 8  8.400413
## 9  8.371612
## 10 8.346330
##
## [[4]]
##                                                          title    score
## 1                                        to kill a mockingbird 9.019460
## 2   the hobbit : the enchanting prelude to the lord of the rings 8.806596
## 3             harry potter and the order of the phoenix (book 5) 8.776832
## 4                                               fahrenheit 451 8.668755
## 5     the fellowship of the ring (the lord of the rings, part 1) 8.627250
## 6                                         silence of the lambs 8.599524
## 7                                     the pillars of the earth 8.570041
## 8  into thin air : a personal account of the mt. everest disaster 8.544208
## 9                                      the secret life of bees 8.505235
## 10                                          bel canto: a novel 8.461241
```

Analysis and Interpretation of Results

Each of the 3 methods gives a different top 10 recommendation for the chosen users however there is some overlap in recommendations from the 3 methods.

We saw that user 23872 and user 63360 are similar with a similarity score of 0.7 and 4 of the same books are in their top 10 recommendations for User Based as expected, namely: – queen of the damned (position 1 for both)
– the hobbit (position 4 and 2 respectively)
– harry potter and the sorcerer's stone (position 2 and 4 respectively)
– the firm (position 7 and 9 respectively)

The item based method also had 3 of these in the top 10: – queen of the damned (position 1 for both)
– the hobbit (position 8 and 3 respectively)
– the firm (position 6 and 9 respectively)
Interestingly, item based recommended books by the same author eg. John Grisham wrote the firm and also the client and the pelican brief from the top 10 recommendations. This makes intuitive sense given that books are considered similar if they have been rated similarly by the same readers, assuming that readers like to read more than 1 book by the same author.

The matrix factorisation method also recommended 2 of the same books for these 2 users. – the hobbit (position 4 and 2 respectively)
– harry potter and the sorcerer's stone (position 2 and 4 respectively)

Therefore, perhaps these 2 books would be appropriate as a combined recommendation from the 3 different methods.

User 11676 has read 148 of the books even though they have only rated 95 so there is only 1 recommendation from items based which is the 1 book not yet read (girl interupted).