Assignment4 Section2 Part1:

Assignment Link:

<https://databricks-prod-cloudfront.cloud.databricks.com/public/4027ec902e239c93eaaa8714f173bcfc/8565400308736738/2355851388225319/1076342497675711/latest.html>

Assignment4 Section2 Part2:

Assignment Link:

<https://databricks-prod-cloudfront.cloud.databricks.com/public/4027ec902e239c93eaaa8714f173bcfc/8565400308736738/2475231989298575/1076342497675711/latest.html>

1) Dataset: Book-Crossing dataset

<http://www2.informatik.uni-freiburg.de/~cziegler/BX/>

2) Examining the dataset:

The Book-Crossing dataset comprises 3 tables.

* BX-Users

Contains the users. Note that user IDs (*`User-ID`*) have been anonymized and map to integers. Demographic data is provided (*`Location`*, *`Age`*) if available. Otherwise, these fields contain *NULL*-values

* BX-Books

Books are identified by their respective ISBN. Invalid ISBNs have already been removed from the dataset. Moreover, some content-based information is given (*`Book-Title`*, *`Book-Author`*, *`Year-Of-Publication`*, *`Publisher`*), obtained from Amazon Web Services. Note that in case of several authors, only the first is provided. URLs linking to cover images are also given, appearing in three different flavours (*`Image-URL-S`*, *`Image-URL-M`*, *`Image-URL-L`*), i.e., small, medium, large. These URLs point to the Amazon web site.

* BX-Book-Ratings

Contains the book rating information. Ratings (*`Book-Rating`*) are either explicit, expressed on a scale from 1-10 (higher values denoting higher appreciation), or implicit, expressed by 0.

Analysis:

As per the given assignment, the relevant csv files are fields are:

**users:**

#User-ID;"Location";"Age"

**bookrating:**

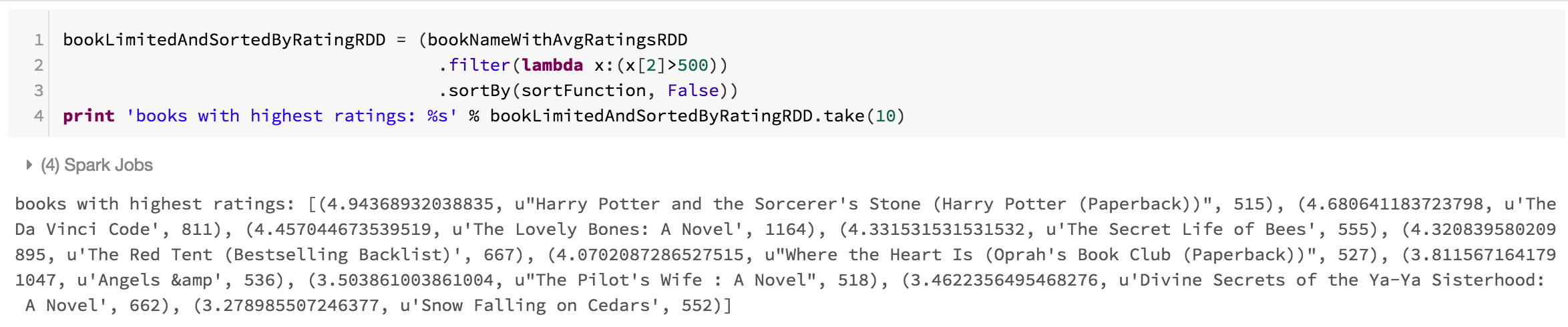
#User-ID;"ISBN";"Book-Rating"

The attributes in the files are separated by semi-colon and some attributes have quotes.

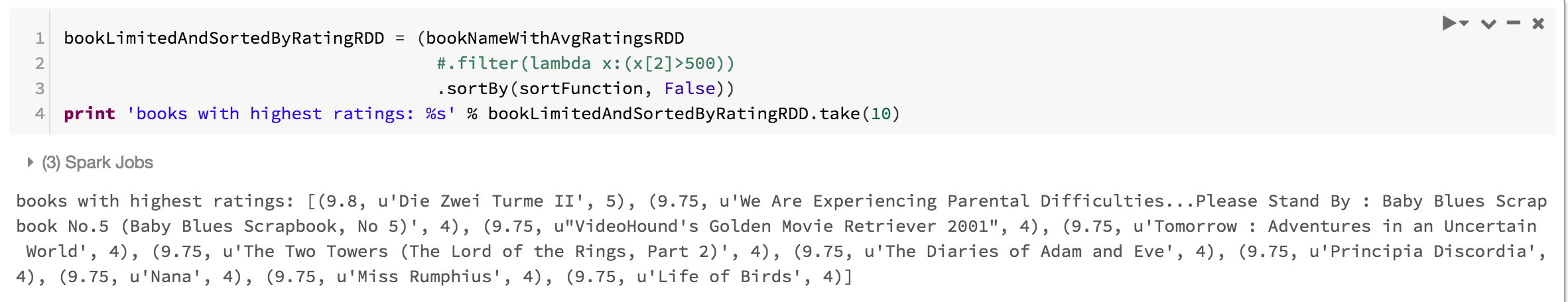


We used. Map() to arrange the data as required.

3) Output: highest average ratings

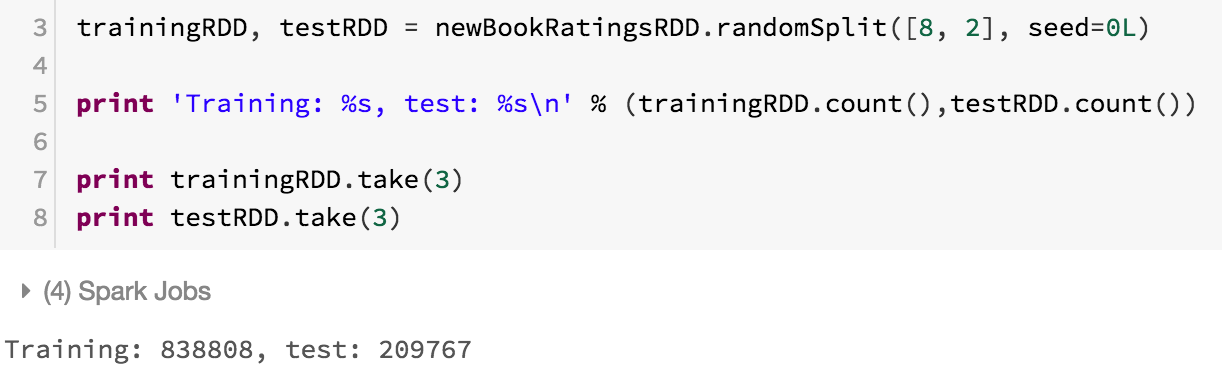


**The highest average rating is for the books which received reviews from more than 500 users.**



**The highest average rating for the books with no condition.**

**4)** Split the data in 80:20 ratio for training and testing.

****

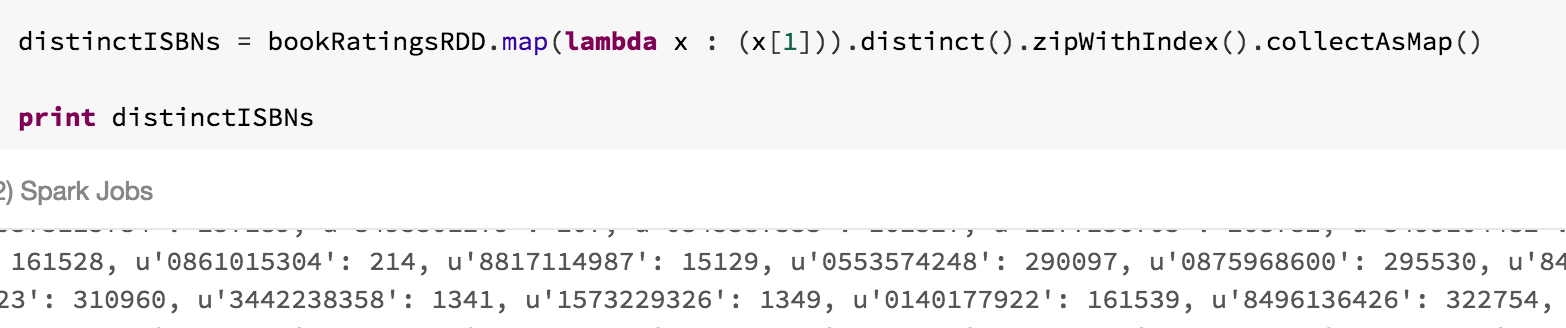
5. Training an ALS model on the training part of the dataset and apply the model on the test part.

Report for RMSE values for different choices of parameters.

ALS Model:

For the ALS model to be created the dataset should be numeric.

Therefore, the strings are replaced by numeric values by creating dictionaries.



Replacing strings with numeric values by using the dictionary created.



|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Seed | Iterations | rank | Lamda\_ | RMSE |
| 1L | 5 | 4 | 0.01 | 7.588 |
| 1L | 5 | 8 | 0.01 | 6.274 |
| 1L | 5 | 12 | 0.01 | 5.801 |
| 10L | 6 | 13 | 0.1 | 4.431 |
| 20L | 6 | 15 | 0.9 | 3.895 |
| 30L | 6 | 17 | 0.99 | 3.884 |
| 30L | 8 | 17 | 0.99 | 3.856 |

6. The URL for the notebook, dataset and name of the dataset

<https://databricks-prod-cloudfront.cloud.databricks.com/public/4027ec902e239c93eaaa8714f173bcfc/8565400308736738/2475231989298575/1076342497675711/latest.html>

Dataset: Book-Crossing dataset

<http://www2.informatik.uni-freiburg.de/~cziegler/BX/>