AYDIN ADNAN MENDERES UNIVERSITY

ENGINEERING FACULTY

COMPUTER SCIENCE ENGINEERING DEPARTMENT

logo, amblem, simge, sembol, ticari marka içeren bir resim

Açıklama otomatik olarak oluşturuldu

**BIG DATA FINAL PROJECT REPORT**

|  |  |
| --- | --- |
| **Project Members:** | **Student Numbers:** |
| Esmanur Deli | 191805056 |
| Kübra Uçar | 191805067 |
| Rabia Yıldırım | 191805043 |

**Project Subject:**

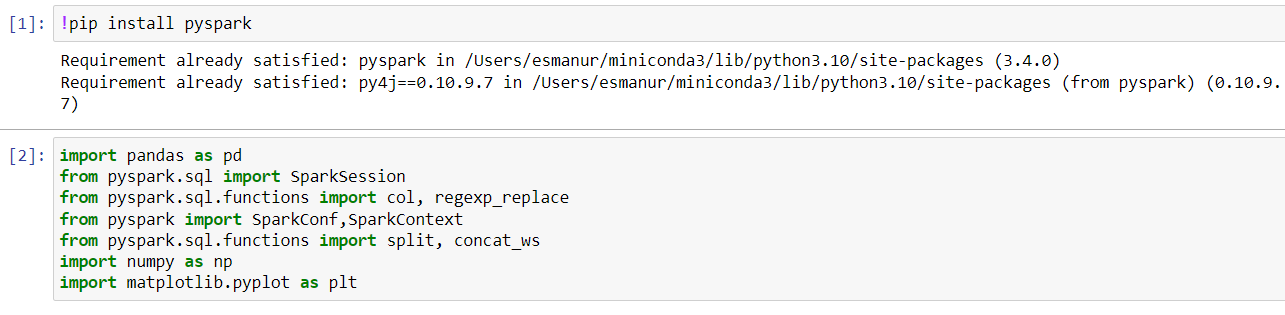
Based on the ratings given to the books by the users and the preferences of the users, the ALS model can make book recommendations and recommend the books to the users that they may be interested in. This way, users can get personalized book recommendations and discover new books.

As a Team, we worked on all parts of project together. You can see the IP Address of computer that we worked on it

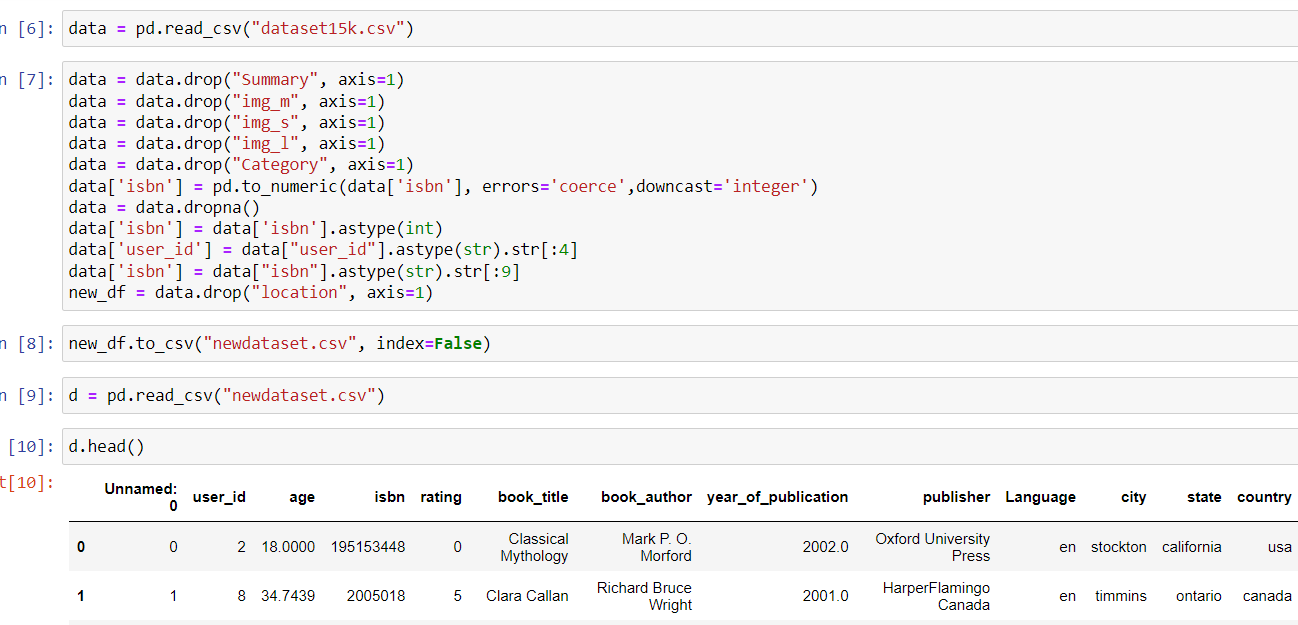
metin, ekran görüntüsü, yazı tipi, çizgi içeren bir resim

Açıklama otomatik olarak oluşturuldu

* Fist of all, we loaded the requirement python libraries and technologies. Then, we imported this libraries. We used Jupyter Notebook as an IDE.



* We found a Book-Rating dataset from Kaggle. This dataset contains 15000 raws and 12 columns as (user id, age, isbn, rating, book title, book author, year of publication, publisher, language, city, state, and country)
* We made some changes on the dataset. We removed some columns we didn't want from the dataset. We changed the data type of some values to integer.
* You can see the some records below from dataset .



* We made the following definitions to perform data processing, analysis or other operations on Spark

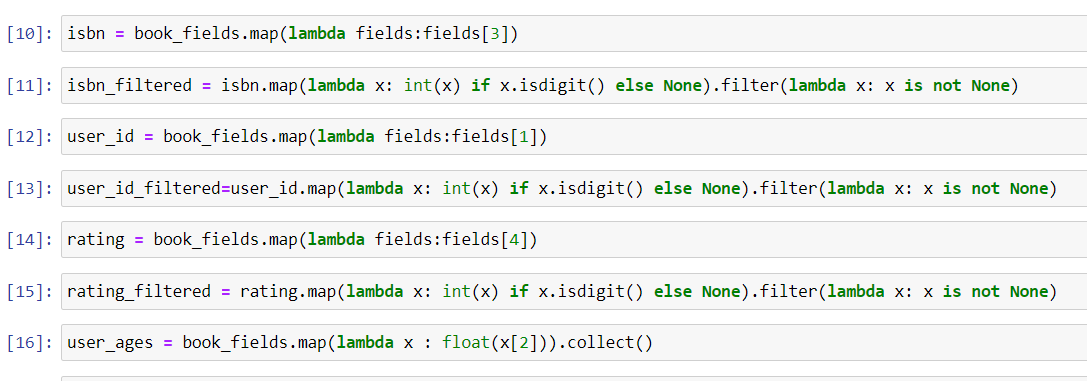


* We took the new csv file and removed the header, then we separated the lines according to the “,” using the map function. You can see the output.

metin, ekran görüntüsü, yazılım, sayı, numara içeren bir resim

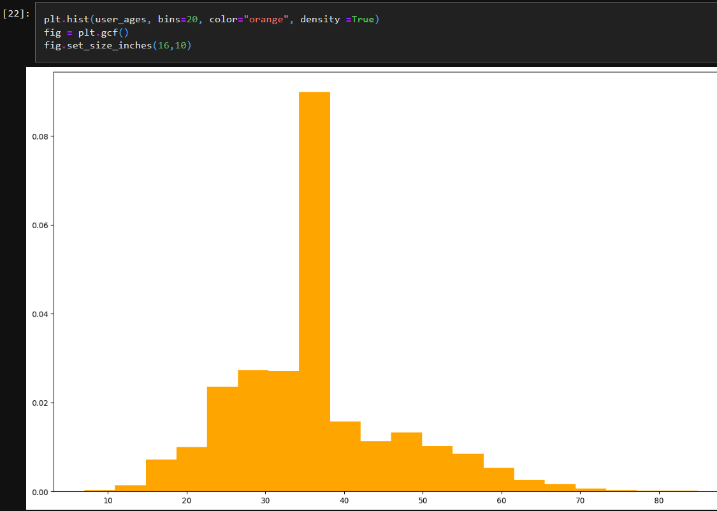
Açıklama otomatik olarak oluşturuldu

* Elements are converted to numeric and values that are none are filtered out.

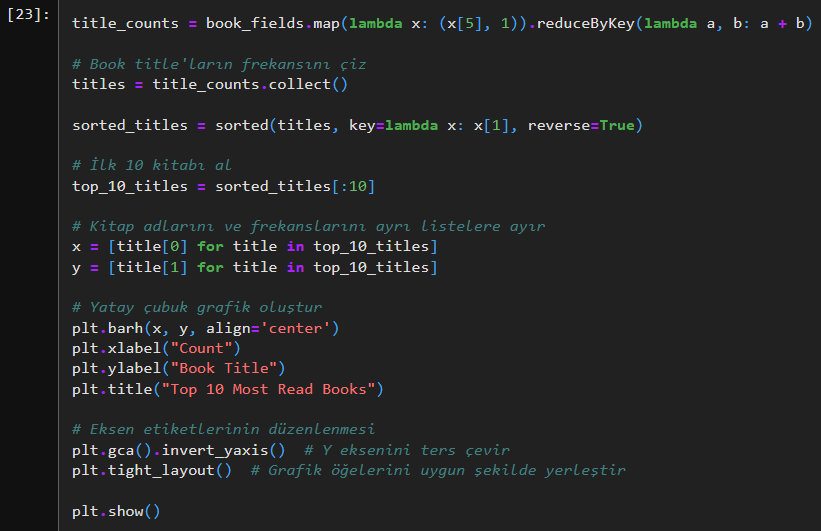


* metin, ekran görüntüsü, ekran, görüntüleme, yazılım içeren bir resim

  Açıklama otomatik olarak oluşturulduHere is a hist chart for user age and rating distribution:



* We have printed the 10 most read books.



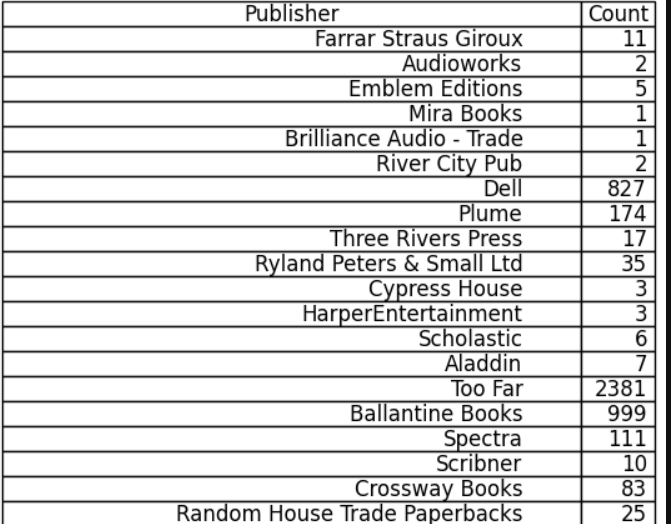
metin, ekran görüntüsü, ekran, görüntüleme, sayı, numara içeren bir resim

Açıklama otomatik olarak oluşturuldu

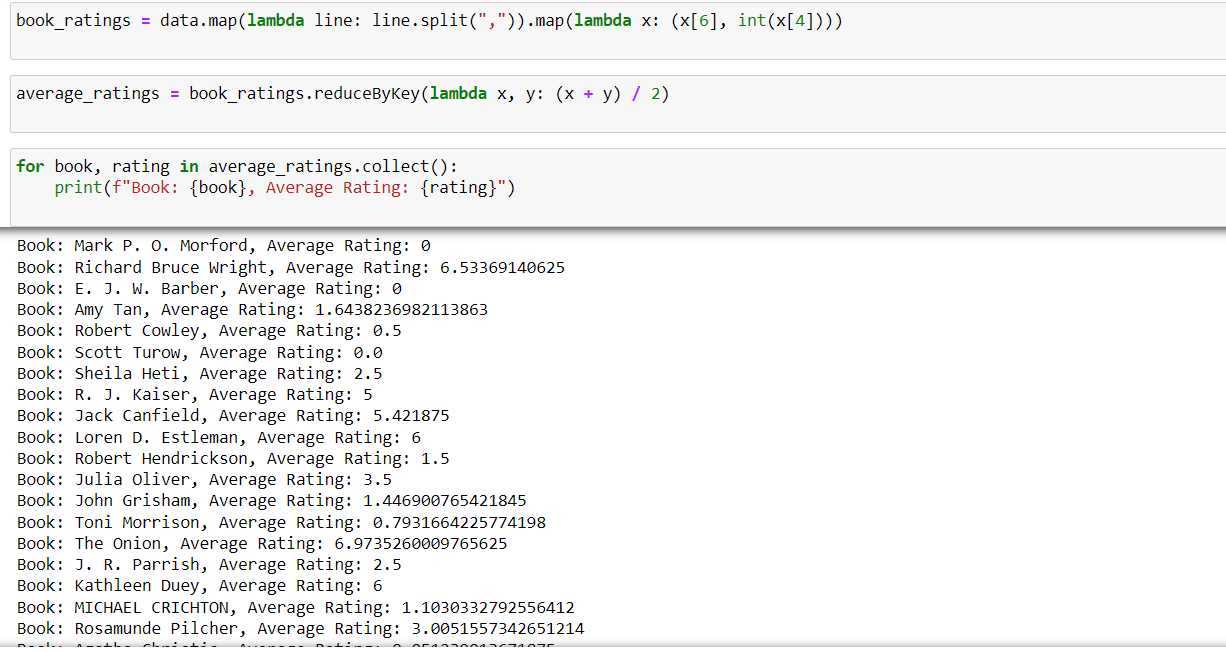
* For publisher distribution, we used map() and reduceByKey() functions together then using collect(), we list these results. So you can see some of how many are from which publishers

metin, yazı tipi, sayı, numara, yazılım içeren bir resim

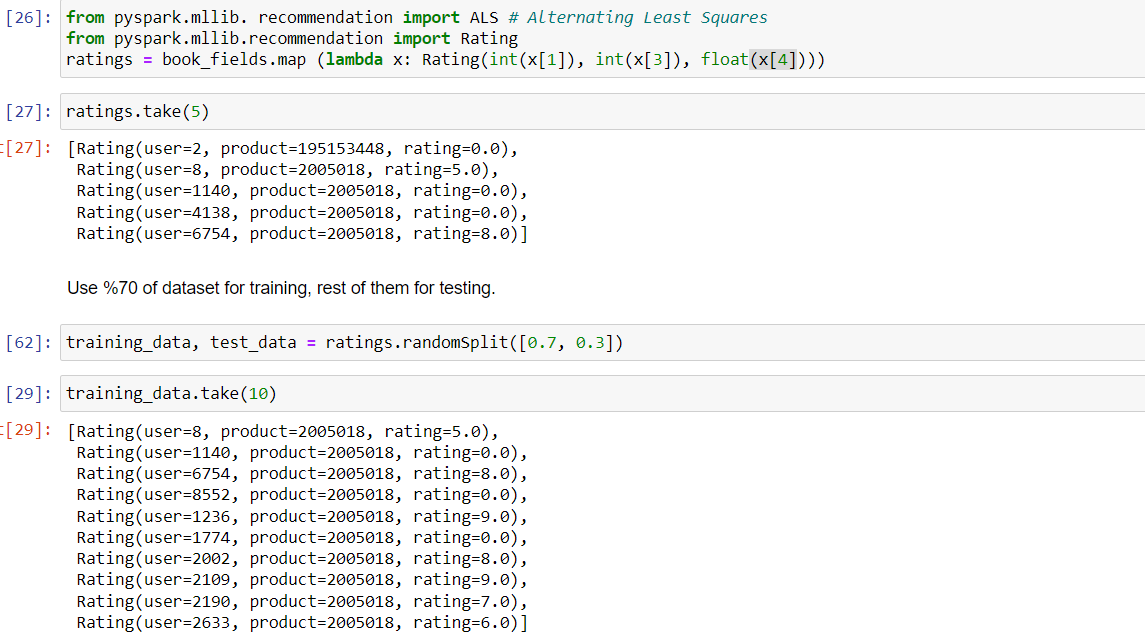
Açıklama otomatik olarak oluşturuldu



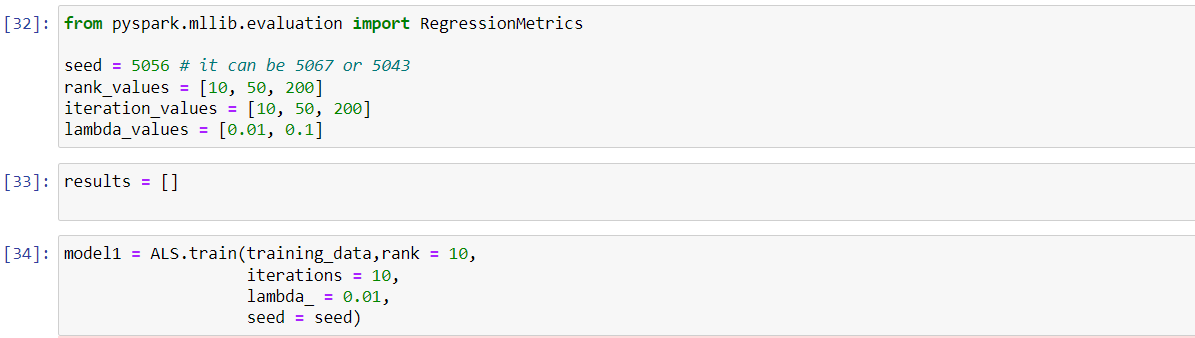
* We have listed the books and the average ratings given by the users for these books, again using the map() and reducebykey() functions.



* Again we imported the libraries for using ALS model. We used “user id, isbn, and rating” columns for recommendation model. Then we split the data as test and train. We made a **70% split for train** and **30% for test**.



* We created a model as you stated in the assignment(seed, “rank” (10, 50, 200), “iteration” (10, 50, 200) and “lambda” (0.01, 0.1):



The following snippet calculates the MSE and RMSE error metrics by comparing the **model's predictions** with the **actual values** in the test dataset to measure the prediction performance of a model.

* We have a problem when we want to do iterations = 200 so we could not create a function to calculate the models with different variables instead we did 12 separate models. You can see example of models below.

metin, ekran görüntüsü, yazı tipi, yazılım içeren bir resim

Açıklama otomatik olarak oluşturuldu

metin, ekran görüntüsü, yazı tipi içeren bir resim

Açıklama otomatik olarak oluşturuldu

* We found the cosine similarity between the ALS model and the produc.

metin, ekran görüntüsü, yazılım, yazı tipi içeren bir resim

Açıklama otomatik olarak oluşturuldu

* We have identified the 10 users who will like the book X the most.

metin, ekran görüntüsü, yazı tipi, sayı, numara içeren bir resim

Açıklama otomatik olarak oluşturuldu