**ADNAN MENDERES UNIVERSITY**

**CSE424 BIG DATA ANALYSIS**

**Term Project**

**Recommendation System with Spark**

**171805024 – Nagihan Baz**

**171805071 – Emin Hasanzade**

**181805052 – Elize Hamitoğlu**

**181805039 – Osman Melih Tolunay**

**Dataset Information**

This dataset is large enough to build good recommendation model and you may face memory issue while training a model using that dataset. Dataset includes books, users and book ratings files.

Books File Description: ISBN, Book-Title, Book-Author, Year-Of-Publication, Publisher, Image-URL-S, Image-URL-M, Image-URL-L.

Book Ratings File Description: User-ID, ISBN, Book-Rating.

Users File Description: User-ID, Location, Age.

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

Açıklama otomatik olarak oluşturuldu

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

Açıklama otomatik olarak oluşturuldu

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

Açıklama otomatik olarak oluşturuldu

Reads files from specified path using textFile with 'sc' method. Returns an RDD where each item represents a row in the file.

Split DF's for each fields



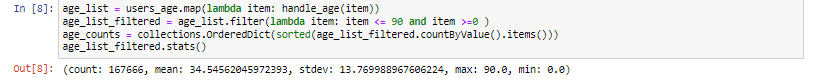
This code creates a new RDD, clearing and transforming the values for further analysis.metin, yazı tipi, ekran görüntüsü içeren bir resim

Açıklama otomatik olarak oluşturuldu

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

Açıklama otomatik olarak oluşturuldu

Calculate baseline statistics for filtered age values.



USERS' Countries Location & TOP 10 Countries That The Most Users Data Collected  
metin, yazı tipi, ekran görüntüsü içeren bir resim

Açıklama otomatik olarak oluşturuldu

Calculates basic statistics for books ratings.



**Computer Information**

metin, yazı tipi, ekran görüntüsü içeren bir resim

Açıklama otomatik olarak oluşturuldu

**MSE (Mean Squared Error)**

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

Açıklama otomatik olarak oluşturuldu

To determine the best model for the dataset based on the Mean Squared Error (MSE) scores provided, we need to identify the model with the lowest MSE score. A lower MSE indicates better accuracy and closer predictions to the actual values.

From the given scores, the model with the lowest MSE is "Model 200\_200\_1" with an MSE of 9.980126049794423. This model has the smallest error, indicating it provides the closest predictions to the actual values in the dataset. Therefore, based on the provided MSE scores, "Model 200\_200\_1" is considered the best model for the dataset. It achieves the lowest MSE, indicating better accuracy in predicting the target variable compared to other models.

**Root Mean Squared Error**

**metin, ekran görüntüsü, yazılım, web sayfası içeren bir resim

Açıklama otomatik olarak oluşturuldu**

To determine the best model for the dataset based on the Root Mean Squared Error (RMSE) scores provided, we need to identify the model with the lowest RMSE score. A lower RMSE indicates better accuracy and closer predictions to the actual values.

From the given scores, the model with the lowest RMSE is "Model 200\_200\_1" with an RMSE of 3.1591337499058856. This model has the smallest error, indicating it provides the closest predictions to the actual values in the dataset. Therefore, based on the provided RMSE scores, "Model 200\_200\_1" is considered the best model for the dataset. It achieves the lowest RMSE, indicating better accuracy in predicting the target variable compared to other models.

**Work Sharing Policy**

Nagihan does the initial research of the project and finds some sample code and different datasets and is responsible for ALS learning and merging the dataset, removing punctuation, removing blank and repetitive items. (Extracted and filtered RDDs.)

Elize created a histogram for the average age, printed the location of the users' countries and the top 10 countries with the most user data collected by code.

Emin was responsible for the cosine similarity function and built the model using the ALS train dataset.

Melih was helpful in researching and writing all the codes in general.

**Dataset Link:** <https://www.kaggle.com/datasets/arashnic/book-recommendation-dataset>

**Note:** The cosine similarity between the ALS model and the products must be found and the 10 users who will like the product X the most should be determined.== We could not do this because we got an error.