SUMMARY

First, we imported the “xml.etree.ElementTree” and “pandas” modules. To work upon the file and to find the correlation coefficient we first converted it to CSV file using “xml.etree.ElementTree” module.

To find the correlation coefficient of the “reputation” of the user in Users.xml to the total answers given by the user in the “PostsHistory.xml“ file, both columns must be integer types then we able to find the correlation coefficient between Reputation PostHistoryId

we use the types function to see all types of columns, and verified that all are of int dtype. We know that the correlation coefficient can be defined as ,

correlation coefficient = covarience(x,y)/std(x)\*std(y)

std =standard deviation

x = Reputation in users.xml

y = PostHistoryId in PostHistroy.xml

In pandas there is already a predefined function “corr” to calculate the correlation coefficient