Relax challenge

Sava Dashev

Relax challenge asks to find all adopted users in data spread into two files - names and engagement.

The names file was encoded using latin-1 and contains symbols that are different than Unicode. Some of the names are not in English. Example: Østergaard Maja

The last session creation time is different format. It needs to be dealt with to change to regular time-date format.

The ratio enrolled for email - not enrolled is 1:3 -roughly 25% of all users.

The different groups are not equal in size.

Process

I added a column for users separating all users that logged more than 3 times. We joined that frame with the engagement frame called data\_all. Further filtering was done to extract only users with 3 or more logins - data\_users.

For these users, I changed the time stamp column into datetime type and extracted the date in a date\_stamp column. All other information from the dataframe data\_users was dumped.

Adopted user

We created list of unique names from data\_users dataframe.

We start with empty list of adopted user.

To find the adopted users, I used a for loop.

We extracted all rows for the current user.

We then run a loop to calculate all differences between the current login and last login. The data is in datetime date format. We subtract the two dates, change to string and extract the substring containing the differences.

In next loop, we calculate sum of two next differences if each of them is greater than zero (ensures different days of login. If the sum of the two differences is less than 8, we add the user to the adopted.

Unfortunately, some dates are not in required format. I deal with them using try-catch block. It makes the script extremely slow.

If I have more time, this will be clean. I do not know how to deal with names not in English, so they stay that way.