Assignment 1:
Bank Customers Similarity

Course: Mining Big Datasets

MSC IN BUSINESS ANALYTICS

PART I – IMPORT DATASET

In this part of the assignment, we loaded the bank customers dataset by using the R programming language and by proceeding to the following processing actions:

- Converted and matched the levels of the *Education* variable (primary, secondary, tertiary) to the ranking levels of 1,2 and 3 respectively.
- Converted the *Job*, *Marital*, *Default*, *Housing* and *Loan* variables to factors.
- Added an extra variable, named *rn*, to uniquely define the ID of each bank customer based on its order within the dataset.

PART II – DATA DISSIMILARITY

In this part of the assignment, a function named *dissimilarityFunction* is created via which the average dissimilarity of any 2 selected bank customers is retrieved. More specifically, the input of this function are the IDs of the desired customers for whom the dissimilarity of **each** attribute (i.e. age, balance, job, marital, education, default, housing, loan) is initially calculated so as for these to be then combined in order to find the average dissimilarity. Below you may see an indicative example for the output retrieved via this function when selecting the bank customers with IDs 1 and 2:

```
> # Call the function to get the average dissimilarity between 2 users
>   dissimilarityFunction(1,2)
[1] 0.393
```

PART III – BANK CUSTOMERS' NEAREST NEIGHBORS

After proceeding to the relevant calculations, the below lists are provided for the top 10 bank customers who present similarity (i.e. the least dissimilarity) with the following bank customers:

Case 1: Bank customer with ID 1230

```
> print(top_10_similar_neigh1)
      UserID AverageDissimilarity
4163    4163                0.000
7208    7208                0.000
35725  35725                0.000
36286  36286                0.000
36607  36607                0.000
37541  37541                0.000
36032  36032                0.001
1906    1906                0.002
2259    2259                0.002
2484    2484                0.002
```

Case 2: Bank customer with ID 5032

```
> print(top_10_similar_neigh2)
      UserID AverageDissimilarity
144      144                0.000
16636  16636                0.000
26741  26741                0.000
30207  30207                0.000
40733  40733                0.000
33843  33843                0.001
380      380                0.002
1775    1775                0.002
6576    6576                0.002
8850    8850                0.002
```

Case 3: Bank customer with ID 10001

```
> print(top_10_similar_neigh3)
      UserID AverageDissimilarity
14250  14250                0.000
16201  16201                0.000
26090  26090                0.000
26784  26784                0.000
35949  35949                0.000
17219  17219                0.001
4317    4317                0.002
10567  10567                0.002
13620  13620                0.002
17229  17229                0.002
```

Case 4: Bank customer with ID 24035

```
> print(top_10_similar_neigh4)
      UserID AverageDissimilarity
9228    9228                    0
10021  10021                    0
13224  13224                    0
17287  17287                    0
18872  18872                    0
19093  19093                    0
19215  19215                    0
20315  20315                    0
20633  20633                    0
20694  20694                    0
```

Case 5: Bank customer with ID 28948

```
> print(top_10_similar_neigh5)
      UserID AverageDissimilarity
1667    1667                0.000
3864    3864                0.000
25686  25686                0.000
30569  30569                0.000
31082  31082                0.000
33068  33068                0.000
35907  35907                0.000
36680  36680                0.000
912      912                0.001
4634    4634                0.001
```

Case 6: Bank customer with ID 35099

```
> print(top_10_similar_neigh6)
      UserID AverageDissimilarity
25245  25245                0.007
30602  30602                0.008
26122  26122                0.009
34720  34720                0.009
1170    1170                0.010
538      538                0.011
2290    2290                0.011
3997    3997                0.011
6040    6040                0.011
7533    7533                0.011
```

Case 7: Bank customer with ID 37693

```
> print(top_10_similar_neigh7)
      UserID AverageDissimilarity
137      137                    0
218      218                    0
1396    1396                    0
2646    2646                    0
6375    6375                    0
6681    6681                    0
7221    7221                    0
7520    7520                    0
10965  10965                    0
14118  14118                    0
>
```

Case 8: Bank customer with ID 39543

```
> print(top_10_similar_neigh8)
      UserID AverageDissimilarity
1604    1604                0.000
4627    4627                0.000
10131  10131                0.000
16297  16297                0.000
26201  26201                0.000
41162  41162                0.000
3421    3421                0.001
4399    4399                0.001
4692    4692                0.001
6464    6464                0.001
```

Case 9: Bank customer with ID 40002

```
> print(top_10_similar_neigh9)
      UserID AverageDissimilarity
28246  28246                0.000
29267  29267                0.000
27203  27203                0.001
40682  40682                0.001
10693  10693                0.002
15202  15202                0.002
27568  27568                0.002
38787  38787                0.002
42053  42053                0.002
43022  43022                0.002
```

Case 10: Bank customer with ID 42192

```
> print(top_10_similar_neigh10)
      UserID AverageDissimilarity
42787  42787                0.000
38868  38868                0.010
22015  22015                0.022
9602    9602                0.023
17648  17648                0.023
41530  41530                0.024
8834    8834                0.025
17475  17475                0.025
20556  20556                0.025
32633  32633                0.026
```

Note that, as can be concluded from the above lists, some bank customers are in fact almost identical with other customers since their dissimilarity (which is rounded in 3 digits) nearly counts to 0.