Sociometer modules documentation –pyspark

The complete workflow involves the usage of the following modules (in the same order).

User profiling

Usage: pyspark user\_profilo.py <folder> <spatial\_division> <region> <timeframe>

Es: pyspark user\_profilo.py dataset centro\_roma.csv roma 06-2015

*Input parameter:*

-folder: the hdfs folder where the dataset is located. In order to let the profiles be computed, it needs at least 3 weeks of data. Dataset is assumed to be splitted into days (e.g. one day = one csv file).

-spatial\_division: A csv file containing the spatial region of each GSM tower. E.g. “RM619D1;city\_center”.

-region: a string containing the name of the region related to the dataset

-timeframe: a string containing the period related to the dataset

*Output:*

It stores the profiles (as Pickle file) into the folder /profiles<region>-<timeframe>.

Profiles are in the format: user\_id->[(region,week n.,workday/weekend, timeframe,number of presence),….]

Clustering

Usage: pyspark clustering.py <region> <timeframe> <archetipi> <k> <percentage>

Es: pyspark clustering.py roma 06-2015 archetipi.csv 100 0.4

*Input parameter:*

-region: a string containing the name of the region related to the dataset

-timeframe: a string containing the period related to the dataset

-archetipi: a csv files containing typical calling profiles for each label. E.g.: Resident->typical resident profiles, etc..

-k: the number of centroids to be computed

-percentage: the percentage of profiles to use for centroids computation

*Output:*

It stores a files “centroids<region>-<timeframe>” containing the association between each centroid and the user type. E.g. Centroid1->resident, etc

User annotation

Usage: pyspark user\_annotation.py <region> <timeframe>

Es: pyspark user\_annotation.py roma 06-2015

*Input parameter:*

-region: a string containing the name of the region related to the dataset

-timeframe: a string containing the period related to the dataset

*Output:*

It stores a file “sociometer<region>-<timeframe>” containing the percentage of user of each profile.

E.g. roma-center, Resident, 0.34