Università di Pisa	A.A. 2015-2016
Data Mining II	

Project assignment

LastFM & Churn

General information

Objective of this project is to perform a few analyses on a dataset of transactions involving the users of the online music service LastFM. The general guidelines for this assignment are the following:

1. the project can be performed by single students or groups up to 3 persons each;
2. each group should perform the tasks / reach the objectives indicated in the text, trying to answer to each request. Any spontaneous addition to that is welcome yet optional, and cannot replace the original TODO list;
3. each group should summarize the work done in a short report (indicatively 5-15 pages), loosely following the guidelines of the CRISP model;
4. each group is totally free to choose the tools and software it prefers;
5. any question, suggestion or request related to the project can be addressed to Mirco Nanni (mirco.nanni@isti.cnr.it).

The dataset

The project will be based on real data describing the listening activities and friendship network of the users of LastFM. The dataset consists of the following tables, provided as CSV files:

listenings.csv : Contains the last 200 listening performed by a set a users. Each listening is characterized by:

- user_id: identifies the user

- date: timestamp of the listening

- track: title of the song listened

- artist: artist of the song

- album: album of the song

genres.csv : Contains an association of the predominant / best fitting genre for a given artist, according to LastFM weights:

- artist: artist

- genre: genre of the artist

network.csv : Contains the network of friendships of the users that have at least a listening in the listenings file:

- user_id1: user_id contained in listening

- user_id2: user which is friend of user_id1 (but not necessarily contained in listening)

Further information on the service can be found at its web address: http://www.last.fm. Also, if needed, LastFM APIs allow to download additional data: http://www.last.fm/api

Objectives

The following activities should be performed and reported:

1. **Exploration**: a **short** data exploration phase, aimed at understanding what data can be useful and whether they present any issues or anomalies.
2. **Artist / Genre churn analysis:** based on the exploration performed above, choose an artist, set of artist or a whole genre and study the churn phenomenon for that, i.e.:
   - Identify the users that consistently used to listen to them
   - Among such users, identify those that, at some point, changed their preferences, and abandoned the artist/group/genre (churn)
   - Study the churn phenomenon, trying to understand what determined it, and build a model able to predict it in advance. The possible causes to consider might include features of the user, of the artist/genre, friends' feaures, etc.
3. **Customer segmentation**: build a customer segmentation of LastFM users based on what they listen to, when they do that, and any other feature you consider relevant.