IST DBDM2 Session 6-7
Practical session on Clustering

Discovering points of interest from geo-localized social media

The city of Lyon (Grand Lyon) wants to improve its massive transportation system. They heard by tracking and monitoring the social networks that a lot of tourists are complaining about the difficulty to reach some interesting sites that the Grand Lyon was not aware of being attracting. Thus, the Grand Lyon decided to use social network data to track the interesting points of interest of Lyon that they may not be aware of. For that matter, they crawled from Flickr (a Yahoo photo sharing system) all the pictures taken and uploaded the last past years in the region of Lyon. At the end of this data collection phase, they have now a database of more that 80,000 pictures.

Each picture can be seen as a tuple

< photo_id, used_id,latitude,longitude,tags, description, dates >

You have been mandated by the Grand Lyon to explore if, yes or no, this dataset can help them to discover points of interest in Lyon. Most importantly, they are interested in knowing which (and how) data mining techniques can be applied. In other words, you need to detail the different phases:

- Clean, describe data (remove duplicates & incoherent values, …)
- Prepare, transform the data (keep informative attributes, select samples, …)
- Mine the data with the clustering techniques (and their parameters)
    - Partitioning with K-means
    - Hierarchical clustering
    - Density-based clustering
- Evaluate and criticize the different results

Useful nodes:

- Pre-processing: *GroupBy, RowFilter, StringToNumber, Geo-Coordinates Row Filter, Missing Values, Column Spliter*
- Mining: *K-means, Hierarchical Clustering, MakeDensityBasedClusterer, Weka Cluster Assigner, NumberToString (PMML), DBSCAN (3.7), Weka Cluster Assigner (3.7)*
- Visualization: *Color Manager, Color Appender, OSM Map View*
- Clustering validation: *Entropy Scorer* if you have a ground truth

Note: Some of these nodes are only available with the full version of KNIME. They can still be installed separately in the basic version by adding the package WEKA and OpenStreetMap