Data Mining : Analyse de médias sociaux géolocalisés

Détection et caractérisation de points d'intérêts

L'objectif de ce TP est d'appliquer des algorithmes de clustering afin de détecter les points d'intérêt (PoI) d'une ville. Une méthode de découverte de motifs (sous-groupes) pourra ensuite être utilisé pour décrire ces PoIs.

Modalité de rendu :

- Taille des groupes : 5 max
- A rendre via Tomuss avant le 30/11, 23:59 :
    - Un rapport sous la forme d'un notebook bien commenté où les choix devront être bien justifiés.
    - Le code associé

Données et prétraitements :

Nous utiliserons des données Flickr sur la région lyonnaise :
https://perso.liris.cnrs.fr/marc.plantevit/ENS/DMTP/flickr_data.csv
Ces données sont très bruitées avec de nombreuses lignes redondantes.

1) Effectuer les prétraitements nécessaires afin de supprimer les doublons et ne retenir que les lignes pertinentes pour l'étude des PoIs à l'échelle de la région lyonnaise.
2) Visualiser les données sur une carte.

Détection de PoIs

Nous souhaitons maintenant détecter les PoIs. Nous définirons un PoI comme une zone dense dans l'espace. C'est-à-dire, une zone sur notre carte avec une forte densité de photos.
Nous souhaitons les détecter en appliquant un algorithme de clustering.

3) Quel type d'algorithme est adapté ?
4) Appliquer l'algorithme(s) retenu(s). Le choix des bons paramètres est déterminant (selon les spécificités des données, on peut imaginer un traitement différent pour certaines zones). On veillera à visualiser les résultats sur une carte.

Caractérisation des PoIs

Une fois les points d'intérêt détectés, nous souhaitons les caractériser. Pour cela, on souhaite exploiter les données, non utilisées pour leur détection, afin de les décrire.

5) A l'aide de règles d'association ou de sous-groupes, caractériser chaque PoI. On veillera à retenir des descriptions qui sont suffisamment discriminantes.

Aller plus loin

6) Quel(s) traitement(s) supplémentaires doit-on faire pour détecter des événements géolocalisés sur ces mêmes données ?
7) Bonus : Détecter et caractériser des événements.

Compléments :

Easy steps to plot geographic data on a map:
https://towardsdatascience.com/easy-steps-to-plot-geographic-data-on-a-map-python-11217859a2db
Clustering: http://scikit-learn.sourceforge.net/stable/modules/clustering.html#clustering
Pysubgroup: https://github.com/flemmerich/pysubgroup
Association rules:

- https://pypi.org/project/apyori/
- http://rasbt.github.io/mlxtend/