# 同济大学软件学院 2019年春季数据分析与数据挖掘课程作业1

## **一．数据集**

本次作业所采用的数据主要基于用户在Yelp（http://yelp.com/）和大众点评(http://dianping.com/)的行为痕迹，包括用户的消费记录、用户对商家的评论以及用户上传的图片。三种数据集的具体格式如下：

1. 用户消费记录数据

* user\_business\_223699.json

{

"uid\_1": ["business\_id\_1", "business\_id\_2", ...],

"uid\_2": ["business\_id\_3", "business\_id\_4", ...],

...

}

该数据集包括了用户的id以及每个用户消费过的商家id。共有223699个不同的用户。

* business\_163665.json

{

"bid\_1": {

"attribute\_1": ...,

"attribute\_2": ...

...

},

"bid\_2": {

"attribute\_1": ...,

"attribute\_2": ...,

...

},

...

}

该数据集包括了 user\_business\_223699.json 中出现过的商家的数据。共有163665个不同的商家。

1. 用户评论数据

* review\_yelp\_ 12992.json

{

"user1\_id": [" word \_1"," word \_2", ...],

"user2\_id": [" word \_1"," word \_2", ...],

...

}

该数据集包括了来自yelp网站的12992位用户的评论数据，其中评论语句为分词提取出的名词形式。

* review\_dianping\_ 12992.json

{

"user1\_id": ["word\_1"," word \_2", ... ],

"user2\_id": [" word \_1"," word \_2", ...],

...

}

该数据集包括了来自大众点评网站的12992位用户的评论数据，其中评论语句为分词提取出名词形式。

1. 用户发表图片数据

图片来自yelp网站，图片数量为64048张，为719位用户去过的商家所上传的图片的总和。user\_photo\_719.json包含了用户以及对应的图片ID。photos\_64048\_user\_719文件夹为所有的图片。

* user\_photo\_719.json

{

"user1\_id": ["photo1\_id"," photo2 \_id", ... ],

"user2\_id": [" photo1 \_id"," photo2 \_id", ...],

...

}

## **二．任务**

1. 阅读并实现论文 《Clustering by fast search and find of density peaks》 。实现所用编程语言不限。 将实现的算法应用在 [Aggregation](http://sse.tongji.edu.cn/zhaoqinpei/Datasets/data/Aggregation.txt)数据集上并将结果绘制在二维平面上，不同的类别用不同的颜色进行区分。
2. 选择以上三种用户数据集中的一个， 以用户为主体进行聚类分析。(a) 用户之间的距离定义是分析过程中一个重要的部分，选择你认为合适的用户间距离定义；在运行聚类算法的时候，需要设置一些参数，其中类的个数是重要的一个参数。(b) 对所选数据集进行分析来确定该数据集的类的个数；(c) 对聚类算法进行实验比较分析（从效率和效果两方面），算法包括 任务1中的算法（sci2014）， KMeans, DBSCAN, Hierarchical, Spectral Clustering 和 EM-GMM 算法；(d) 选择合适的评价指标对不同算法的聚类结果进行评估， 并针对每一种算法记录最佳的聚类结果；(e) 综合以上几个方面，分析结果并写成报告。

**注：**

1. **本次作业涉及到的所有数据集以及论文均可以在**[**http://10.60.43.58:9579/hw1\_data/**](http://10.60.43.58:9579/hw1_data/) **上进行下载。**
2. **提供的用户数据集是随机选取的，可根据实际情况进行筛选和处理，实验中不一定要用到全量数据。除Task 1以外，其他算法可以调用现有的实现，不会影响评分。**
3. **DBSCAN的参数设置要求聚类结果中类的个数接近类的个数的分析结果，i.e., 等于或接近其它聚类算法中类的个数设置。**

## **三．提交**

提交日期：2019-4-19 23:59:59，提交至Piazza。提交内容要求：

提交文件命名为学号\_姓名(中文)\_hw1.zip。共有两个子目录，对应两个任务，命名为q1，q2，每个子目录包括以下内容：

1) 源代码文件。

2) README文件，介绍运行环境和运行方式。

3) 实验报告文件，包括数据预处理、论文实现过程中的亮点和难点、用户间距离定义，实验结果以及对实验结果的比较分析等。

4) 实验结果文件。任务1 和 任务 2 各一个文件，均为csv文件格式。每行具体内容如下：

Task 1: [ x, y, label ]

Task 2: [uid, sci2014\_label, kmeans\_label, dbscan\_label, hierarchical\_label, spectral\_label, em\_label ]. e.g. [bIzoX\_6PNnpxPiwhJeUMfg, 0, 1, 1, 7, 12, 8 ]