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。

2. 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。

3. 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 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/ 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]