- 人群聚类
  - 简介
    - 通过对每个用户提取不同的属性，例如年龄，城市，爱好，等数十个标签，做为不同的特征，输入到sklearn的聚类模型
  - 项目流程
    - 1. 爬取各类型数据
    - 2. 数据处理，人群的各类型数据分配不同的属性和属性标签
    - 3. 分析师选定感兴趣的属性
    - 4. 根据感兴趣的属性作为特征，输入到sklearn，返回人群分类结果
    - 5. 对输入结果进行分析，分配类别
  - 代码流程
    - 示例项目：gucci项目
    - 示例代码：my-da/work/run_0824.py
      - 函数extract_user_data_1： 根据excel列表匹配感兴趣的字段
        - 每个任务需要修改较多，主要是进行匹配的列的名字
          - 例如输入：gucci/old/gucci_example_1000.xlsx
          - 例如输出：gucci/input/user_data_all.xlsx
      - 函数select_data： 过滤数据，过滤一些空数据
        - 过滤掉收集到的不需要的行数据，例如空数据
          - 输出：gucci/input/m_user_feature.xlsx
      - 函数do_cluster：聚类，调用sklearn进行聚类
        - 特征输入到模型聚类
          - 输出：gucci/output/m_group_all_result.xlsx
      - 函数extract_user_group_info：整合聚类结果，例如TGI信息生成
        - 根据聚类结果，结合输入的excel进行生成分析结果
          - 输出：gucci/output/last_group_result.xlsx
  - 目录结构
    - 数据目录
      - 项目名字的目录下有3个子目录，分布是old，input，output
        - old是保存excel源数据
          - gucci_example_1000.xlsx
        - input是保存属性筛选结果
          - user_group_merge.xlsx
          - m_user_feature.xlsx
          - user_data_all.xlsx
        - output是保存聚类和整个结果
          - last_group_result.xlsx
          - m_group_all_result.xlsx
    - 代码目录
      - 码，主要修改这里

```
├── fastal  算法库
│   ├── algorithms
│   ├── analysis
│   ├── app
│   ├── bus
│   ├── common
│   ├── ml
│   ├── nlp
│   └── util
└── my-da
    ├── config  配置文件
    └── work. 独立的项目代
```

  - 任务的配置信息
    - config.py文件
      - 包含字段： 项目名称
      - 聚类类别数量
      - 特征
        - 名字是：聚类时结果显示的列名，例如：品牌， 即聚合是的名称，对应不同的聚合的值
        - 包含多个匹配
          - 匹配excel列
            - 即源数据的excel列
          - 匹配词
            - 可以为None，string或list
            - None表示所有词都可以
            - string是只匹配这个词，必须匹配excel列的内容和这个string相等
            - list表示这个匹配excel列的内容在这个列表中，就算匹配上了
          - 不匹配词
            - 相对于匹配词，如果出现不匹配词，那么也进行筛选
          - 聚合的值
            - 如果聚合的值存在就按聚合值作为值，否则就按匹配excel中的内容作为聚合值
        - 如果存在level3字段
          - 那么说明需要进一步匹配，进一匹配excel的列
            - 匹配excel列
              - 即源数据的excel列
            - 如果有level4
              - 列表格式，每个item包含匹配词和聚合值
              - 当匹配excel列的内容在匹配词列表中的时候，聚合的值就是聚合值字段
            - 如果没有level4
  - 项目的问题
    - 聚类绘图
      - Excel中的关注的所有特征数量绘图
      - 匹配到的所有特征数量分布绘图
      - 人群聚类结果绘图（人群数量和类别bar图）
      - 特征经过PCA降维的2维后的散点图
      - 特征经过PCA降维的3D散点图
    - 聚类后某些特征分不开，例如聚类选择了30个特征，有的特征全部属于一个类别
    - 有的时候需要对已经聚类过的人群出来的类别，当部分数据更新后，仍然聚类出原来的类别，目的是看这些类别的人群经过一段时间后的改变