数据标注说明

1. 现有数据

（1）原始数据

从 JD 采集的农产品属性和评论数据。

（2）产品品类

28 个品类：稻谷、花生、玉米、棉花、芝麻、洋芋、小麦、大豆、番茄、蜜茄、脚板薯、藕、竹荪、竹笋、大米、大蒜、姜、豆皮、橙皮、茶叶、李子、荸荠、椪柑、梨、枇杷、猕猴桃、蜜柚、茶油。

（2）评价方面

8 个方面：价格、品质、色泽、口感、包装、分量、物流、售后。

主题词典：每个方面对应的关键词集合。

方面	关键词
口感	酱料，味道，甜味，气味，甜度，酸味，…
物流	取货，效率，仓储，收货，提货，进货，…
品质	水准，废物，质量，高品质，低质量，品位，…
价格	让利，促销，不值，降价，总价，不屑，…
色泽	色调，紫色，棕色，花纹，蓝绿色，色泽，…
分量	颗粒，体积，大小不一，微小，微粒，块头，…
包装	包装袋，皮箱，包装，生锈，麻袋，渗漏，…
服务	热情服务，客户服务，售后服务，客服热线，…

2. 标注流程

（1）数据采样

按照（28 个）品类×（8 个）方面×（5 档）评分将评论数据划分为 28×8×5=1120 个子集，从每个子集中随机抽取 10 条评论用于标注（共计 11200 条）。

补充说明：

① 评论文本长度限制在 2～40 字；

② 根据主题词典大致确定评论涉及的方面；

③ 若子集中评论数量不足 10 条，则全部抽取；

④ 确保抽取出的评论文本不存在重复。

（2）数据标注

待标注的数据已包含：产品品类、评论文本、用户评分。

标注评论文本中的评价单元，以四元组的形式定义：<评价对象，观点表达，评价方面，情感极性>。

因此，需要标注的内容包括：评论文本中的评价对象，以及每个评价对象对应的观点表达、评价方面、情感极性，具体定义如下：

评价对象指的是用户在评论中提及的实体（如花生、玉米、大豆…）或属性（如质量、价格、味道…）。这里不对实体和属性进行区分，统称为评价对象。

例：果仁不大还有糊的，体验较差。

<果仁>

<体验>

观点表达指的是用户在评论某个评价对象时具体的语言表达，一般为副词、形容词、动词、短语等。

例：果仁不大还有糊的，体验较差。

<果仁，不大>

<果仁，有糊的>

<体验，较差>

评价方面指的是评价对象所属的方面类别。若评价方面不属于上述 8 个方面中的任意一个，则归入“其他”方面。因此，实际上共有 8+1=9 个方面类别。

例：果仁不大还有糊的，体验较差。

<果仁，不大，分量>

<果仁，有糊的，色泽>

<体验，较差，其他>

情感极性指的是用户对某个评价对象的情感态度，分为正面、负面和中性。

例：果仁不大还有糊的，体验较差。

<果仁，不大，分量，负面>

<果仁，有糊的，色泽，负面>

<体验，较差，其他，负面>

补充说明：

① 评价对象和观点表达需要记录它们在评论文本中的位置，以起始位置索引和结束位置索引表示。

例：果仁不大还有糊的，体验较差。

<果仁[0, 1]，不大[2, 3]，分量，负面>

<果仁[0, 1]，有糊的[5, 7]，色泽，负面>

<体验[9, 10]，较差[11, 12]，其他，负面>

② 评价对象（字符串）和观点表达（字符串）可以缺省，评价方面（9 选 1）和情感极性（3 选 1）不能缺省。

例：难吃死了。

<--，难吃死了[0, 3]，口感，负面>

（3）数据存储

完成标注的数据以 JSON 格式存储在文件中，文件每行是一个字典，记录一条评论及其标注信息，各字段定义如下：

```
{
    “comment_text”: xxx,  // 评论文本
    “comment_variety”: xxx,  // 产品类型
    “user_star”: xxx,  // 用户评分
    “items”: [  // 评价单元列表
        {
            “target_text”: xxx,  // 评价对象
            “target_index”: [x, x],  // 评价对象的起始位置
            “opinion_text”: xxx,  // 观点表达
            “opinion_index”: [x, x],  // 观点表达的起始位置
            “aspect”: xxx,  // 评价方面
            “sentiment”: xxx,  // 情感极性
        },
```

…

]

}

3. 其他说明

在评价单元的标注过程中，需要记录评价对象和观点表达的起始位置和它们之间的对应关系。为了支持上述功能，简化标注过程，可能需要制作一个前端界面辅助标注工作。一种可能的界面设计方式如下：