# 周报

屈原斌

2020 年 10 月 5 日

## 1 上周任务

- 数据处理；

### 1.1 数据处理

#### 1.1.1 数据筛选

- 主要筛选四类卷（390个题目）和五类卷（237个题目），最终保留193个题目，共1930篇作文；
- 方法：
  - 题目筛选：
    - 去除要求字数¡600的题目，共**123**个题目；
    - 去除多主题的题目，共**4**个题目；
    - 去除重复题目，共**7**个题目；
  - 作文抽取：
    - 在剩余题目中筛选出四类卷和五类卷重叠题目且题目下作文数**>10**的题目，共**144**个题目，每个题目抽取**10**篇作文（保证既有四类卷作文，又有五类卷作文）；
    - 在去除重叠题目后的四类卷题目中筛选作文数**>10**的题目，共**49**个题目，每个题目抽取**10**篇作文；
- 结果：
  - 最终保留了四类卷和五类卷共**193**个题目，共抽取**1930**篇作文；

### 1.1.2 数据标注

- 标注规范（见表1），共分为4档评分；
- 抽取部分数据进行标注：
  - 四类卷作文（标注结果见表2）：
    - **数据：** 共抽取7个主题（5个命题作文，1个半命题作文，1个自拟题目），共320篇作文；
    - **结论：** score=3的作文占比最多，完全离题作文较少；
    - **问题：** 作文转写质量较差，可能会影响标注结果；
    - score=4部分分数低的原因分析：
      - 结构混乱，逻辑上不通顺；
      - 篇章结构不完整，没有写完；
  - 三类卷作文（标注结果见表3）：
    - **数据：** 抽取2个主题，每个主题抽取10篇作文；
    - **结论：** 全部为不离题作文；
  - 五类卷作文（标注结果见表4）：
    - **数据：** 抽取2个主题，每个主题抽取10篇作文；
    - **结论：** 离题作文占比较大，内容重复较多（摘抄材料）；

表 1: 离题标注规范

| **score** | **标准** |
|---|---|
| 4 | 不离题，文章内容始终符合题目要求； |
| 3 | 部分离题，内容基本符合题目要求，偶尔有离题部分 |
| 2 | 部分离题，部分内容符合题目要求（离题部分超过整篇文章50%以上） |
| 1 | 完全离题，文章内容和题目没有关系（包括恶意提交、流水账作文等） |

表 2: 四类卷标注结果

| 题目 | score=1 | score=2 | score=3 | score=4 | 作文数 |
|---|---|---|---|---|---|
| 你是我的一面镜子 | 0 | 9 | 11 | 33 | 53 |
| 你，就这样留在了我的记忆里 | 1 | 2 | 8 | 43 | 54 |
| 捡拾幸福 | 2 | 5 | 18 | 22 | 47 |
| 永远的风景 | 2 | 2 | 6 | 30 | 40 |
| 新结识的朋友 | 0 | 4 | 10 | 41 | 55 |
| __的足迹 | 0 | 2 | 2 | 36 | 40 |
| 为父亲或母亲写一篇小传(自拟题目) | 0 | 0 | 0 | 31 | 31 |
| **总计** | **5** | **24** | **55** | **236** | **320** |

表 3: 三类卷标注结果

| 题目 | score=1 | score=2 | score=3 | score=4 | 作文数 |
|---|---|---|---|---|---|
| 我有一颗__的心 | 0 | 0 | 0 | 10 | 10 |
| 我把掌声送给您 | 0 | 0 | 0 | 10 | 10 |
| **总计** | **0** | **0** | **0** | **20** | **20** |

表 4: 五类卷标注结果

| 题目 | score=1 | score=2 | score=3 | score=4 | 作文数 |
|---|---|---|---|---|---|
| 最是难忘那表情 | 5 | 4 | 1 | 0 | 10 |
| 记住这一天 | 3 | 2 | 0 | 5 | 10 |
| **总计** | **8** | **6** | **1** | **5** | **20** |

## 2 本周计划

- 数据对接，开始数据的标注工作；
- 复现一些的论文中离题检测的方法；