**表头识别分析**

**1.识别规则**

（1）取表格前5行

（2）统计每一行词语在词典中的出现比例，设匹配的个数为，该行的词语个数。则比例计算为：

（3）计算每行词语出现在统计词典中的词语频率，前5行词语在词典中的出现总频率，则频率比例为：

（4）总公式为：

（5）设定每一行的匹配词语数要大于等于2.

（6）设定阈值，小于阈值的则标为-1（即表头不在前5行）.大于等于阈值，且该行匹配词语数大于等于2的，则取的最大值的所在行作为表头。

**2.数据分析**

（1）label：

|  |  |
| --- | --- |
| **Label** | **Description** |
| -1 | 表头不在前5行 |
| 0 | 表头在第1行 |
| 1 | 表头在第2行 |
| 2 | 表头在第3行 |
| 3 | 表头在第4行 |
| 4 | 表头在第5行 |

（2）数据统计

|  |  |
| --- | --- |
| **Label** | **Number** |
| -1 | 26 |
| 0 | 1 |
| 1 | 11 |
| 2 | 6 |
| 3 | 7 |
| 4 | 1 |
| **Precision:1.0 Recall: 1.0** | |

**3.真实场景统计**

总共有24张表，去掉两张不是表格的数据，剩下22张表

|  |  |
| --- | --- |
| **Label** | **Number** |
| -1 | 0 |
| 0 | 6 |
| 1 | 10 |
| 2 | 4 |
| 3 | 2 |
| 4 | 0 |
| **Precision:1.0 Recall: 0.954545** | |

1.经过处理之后，去掉了不是表格数据的《中小学生参加校外培训调查表》和《学生毕业生登记表》。在剩下的22张表中，得到的结果为precision: 100%和recall: 95.5%。

2.没有召回1个case

错误的表格：《【无标题】合同进展每周反馈》；处理结果：-1；期望结果：0