**处理前**

|  |  |  |  |
| --- | --- | --- | --- |
| 分类 | 样本数量 | Text长度 | Text区间 |
| 法律 | 31k | 124 | 8~4334 |
| 各部门规定 | 85k | 132.7 | 8~ 32767 |
| 条例 | 22k | 117.9 | 9~5486 |
| 地方法 | 557k | 125.8 | 4~ 11397 |
| 地方规定 | 250k | 121.8 | 4~32763 |
| 企业文化 | 2477 | 252.2 | 17~16236 |
| 道德基础 | 656 | 97.7 | 29~251 |
| 总计 | 945k |  |  |

**处理后**

处理方式：剔除超过1000字符的text

|  |  |  |  |
| --- | --- | --- | --- |
| 分类 | 样本数量 | Text长度 | Text区间 |
| 法律 | 31k | 122.8 | 8~952 |
| 各部门规定 | 85k | 120.5 | 8~999 |
| 条例 | 22k | 117.2 | 9~981 |
| 地方法 | 557k | 125.2 | 4~1000 |
| 地方规定 | 250k | 119.3 | 4~999 |
| 企业文化 | 2477 | 227.1 | 17~990 |
| 道德基础 | 656 | 97.7 | 29~251 |
| 总计 | 945k |  |  |

设置上下限20~1000

|  |  |  |  |
| --- | --- | --- | --- |
| 分类 | 样本数量 | Text长度 | Text区间 |
| 法律 | 31k | 123.2 | 20~952 |
| 各部门规定 | 85k | 121.3 | 20~999 |
| 条例 | 22k | 118.2 | 20~981 |
| 地方法 | 557k | 125.6 | 20~1000 |
| 地方规定 | 250k | 119.7 | 20~999 |
| 企业文化 | 2477 | 227.4 | 23~990 |
| 道德基础 | 656 | 97.7 | 29~251 |
| 总计 | 945k |  |  |