**数据脱敏流程**

全部的脱敏过程是使用python代码本地处理excel表格中的信息。

1. **去除电话号码：**

文本信息中除去空格、横线(-)、下划线后(\_)检测文本中的连续数字，若文本中包含连续8位及以上的数字，则用\*隐藏相应的数字串，

例如：

**原信息：**我今年30岁，我的电话号码是13911122233，我的qq号是1234512345.

**处理后：**我今年30岁，我的电话号码是\*\*\*\*\*\*\*\*\*\*\*，我的qq号是\*\*\*\*\*\*\*\*\*\*.

1. **去除邮箱：**

文本信息中检测是否存在”@”和”.”且“@”在“.”之前的格式，若检测到这样的格式，则用”\*\*\*\*”替换”@”前的全部内容，以及”@”和”.”之间的全部内容。

例如：

**原信息：**[someone@lenovo.com](mailto:someone@lenovo.com)

**处理后：**[\*\*\*\*@\*\*\*\*.com](mailto:****@****.com)

1. **去除地址：**

该项使用了python cpca库，是一个开源的提取中文省市区详细地址的库，使用该库后可以提取文本信息中的地址信息为Python中的pandas DataFrame结构体，结构如下：

省 市 区 地址

0 北京市 北京市 海淀区 后场村2路461号58号楼5楼

提取完成后回到原文中寻找对应信息，将提取出来的省信息替换为”\*\*省”，市信息替换为”\*\*市”，区信息替换为”\*\*区”，地址替换为”\*\*\*\*\*”.

Python cpca库参考资料：<https://blog.csdn.net/xc_zhou/article/details/107169506>

例如：

**原信息：**我在浙江省杭州市的金地广场

**处理后：**我在\*\*省\*\*市\*\*区\*\*\*\*\*

1. **去除姓名：**

该项使用了python hanlp库，是一个开源的提取中文姓名的库，使用后可以将文本中所有疑似姓名的字段提取出来成列表格式输出。在提取出姓名后，把原文中对应的部分替换为”\*\*\*”。

Hanlp参考资料：<https://github.com/hankcs/pyhanlp>

例如：

**原信息：**您好，我是百应工程师徐某某，很高兴为您服务。

**处理后：**您好，我是百应工程师\*\*\*，很高兴为您服务。