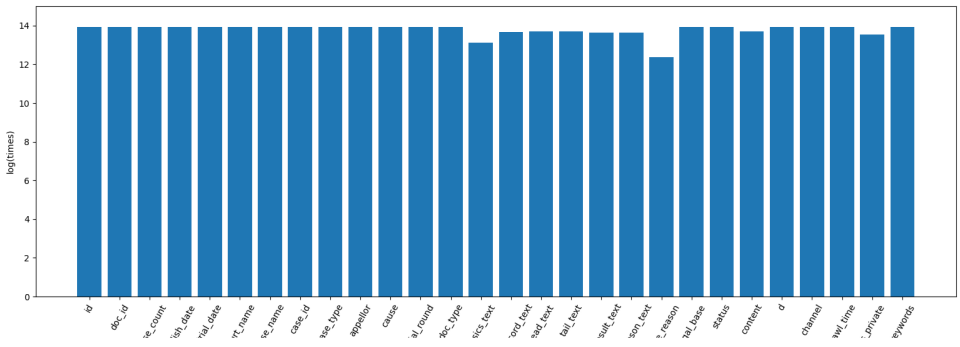
数据预处理

原给定的数据集共有1125799个记录，每个记录有50个数据列，其中有个别数据列全为空。



以content列为依据， 统计发现，content字段为空时，其他text字段也几乎为空。

所以目前处理的步骤是： 删除字典里面全部为空的字段，并且删除content字段为空的记录。

然后删除文本中多余的格式控制符号。

目前共有88w余条数据保留。