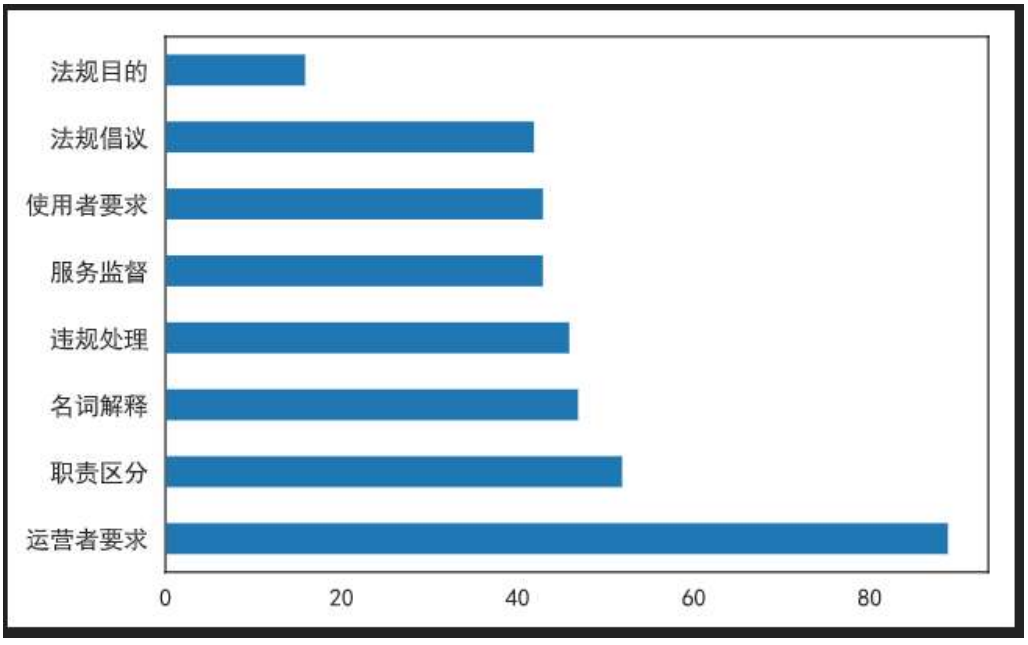
发现

1.XGBClassifier

注意到tf-idf文件中的XGBClassifier分类器模型，其基准训练准确率是 `0.702` ，效果较好，进一步观察混淆矩阵，发现存在大量的各类标签被错误归入 运营者要求 类别中，特别是以 使用者要求 、 服务监督 、 法规倡议 为甚。但 运营者要求 类别的数据却较少归到其它标签中（除了0.15的 使用者要求 ） 这说明目前 运营者要求 的接纳性比较大，很容易符合其它标签数据的阈值，进而错误划分。

2.数据占比

从下图可以看见， 运营者要求 的数量最大，这有可能是造成 运营者要求 的接纳性大的一种原因，后续增加训练数据时，不要再增加 运营者要求 数据，平衡一下各类别数据，或许效果会更好。



3.原始数据

仔细观察原始训练集的数据可以看到， 运营者要求 和 使用者要求 两项确实很容易雷同，它们在逻辑上的划分是， 运营者要求 是法规对于互联网信息服务的提供商（即企业）所提出的要求，多半后与技术要求、政府职能对接， 使用者要求 是法规对于所有上网的人提出的要求，多半后与一些比较大的词汇，或者禁止词汇挂钩。所以如果要有比较高的区分能力，方法最好能把两种标签类别的主语给区分开，不然的话这两种的查准率就不会高。

建议（供参考）

1. 在增加数据集的时候，着重增加 使用者要求 、 服务监督 、 法规倡议 三种类别的数据，或者说在遴选数据的时候，增加该三者的比重，使它们在分类的时候不会因样本不足而被 运营者要求 错归。
2. 重新排查一遍 运营者要求 与 使用者要求 的已分类数据，更正里面错归的个例。
3. 尝试将 运营者要求 与 使用者要求 暂时先归为一类，并使用tf-idf方法、xgboost进行测试，看看两类归一后会不会有较大的性能提升。
4. 从二元组的报告看，我感到二元组理论上可以很明显地区分 运营者要求 与 使用者要求 ，因为它们的主语与后接对象往往是不一样的搭配，但是实际的测试效果并不好，能否做一个单独的测试，查看二元组的办法（包括词频使用tf-idf加权后的二元组）能否对 运营者要求 与 使用者要求 的 **二分类问题**起效果。