项目说明

数据划分

将数据按照9比1的比例分为训练集与测试集。由于数据分布不均匀，在数据处理中将训练集与测试集中每个分类的占比都保持一致。

特征提取

由于用户浏览url的序列具备前后文关系，故采用word2vec的模型对内容做向量化处理。在处理的时候，将每个url也打散成单词（分割符包括"/"、"-"等等），这样处理后的准确率经实验验证比之间使用原始**url**效果要好一些，猜测也许url中的层级关系中也包含某些语义成分。

模型选择

由于存在标签数据分布不平衡的情况，故采用可以规避此类问题的随机森林、AdaBoost、Gradient Tree Boosting三种模型进行实验比较。以上三种方法由于混合了多种模型，不会产生过拟合问题，所以可以不使用交叉验证集。

最后为了利用序列信息，尝试了一下lstm的模型。

最后测试结果

结果看下来，GradientBoosting和lstm效果都还算不错

Random Forest
模型准确率 70.939490446%
purchaser recall:32.01754386%
purchaser precision:40.782122905%
supporter recall:87.823439878%
supporter precision:87.556904401%
researcher recall:64.959568733%
researcher precision:59.359605911%

Adaboost
模型准确率 71.576433121%

purchaser recall:28.50877193%
purchaser precision:38.922155689%
supporter recall:88.432267884%
supporter precision:88.432267884%
researcher recall:68.194070081%
researcher precision:60.238095238%

Gradient Boosting
模型准确率 73.8853503%
purchaser recall:32.894736842%
purchaser precision:51.020408163%
supporter recall:88.127853881%
supporter precision:88.396946565%
researcher recall:73.854447439%
researcher precision:61.990950226%

LSTM
模型准确率 74.919614148%
purchaser recall:28.947368421%
purchaser precision:51.162790698%
supporter recall:89.802130898%
supporter precision:87.407407407%
researcher recall:73.315363881%
researcher precision:62.385321101%

可能再改进的方向

由于数据不平衡，可以尝试再手工制造一些数据，比如oversampling等方案，也许会有好的效果。