web安全

拼音与英文识别实现

实现思路一：基于字符频率的朴素贝叶斯分类法。通过英文字母表频率与拼音表频率，根据某字符串的字母组合可以计算出某字符串对两类别的概率，进而实现分类。但是该方法对较短的字符串识别效果不好，只有字符串足够长的文段才适合基于字符频率的分类识别。而且该方法忽略字符间的位置信息，识别效果不够理想。

实现思路二：基于深度学习实现拼音与英文识别。该方法把不同长度的字符串padding为统一长度，通过构造LSTM模型进行分类识别。该算法需要一定量的已经标注过的数据集才能训练，且消耗计算资源更多，开销大。

实现思路三：模板匹配。人的大脑对字符组合储存有限，其实不管是英文还是拼音，设计的组合都是有限的。可以通过对英文数据库和拼音数据库进行比对，对字符串进行精准识别。英文数据库内含十万多的英文单词，拼音数据库有四百多种不同的拼音组合。直接字符串匹配速率较慢，因此我提出了一个26叉字母树的存储结构用于存储10万个单词数据库的英文字母表，避免每个字符串需要每次匹配10万次的计算量，实现更高速率的匹配。

最终选择思路三进行实现，该算法在准确率和匹配速率上都表现出了较好的结果。