数据处理中只保留中文发现的问题：

1. 训练集第32621行，搜索内容全部为百度链接，处理后该行内容为空

2. 训练集第32621行，搜索内容全部为英文单词，处理后该行内容为空

解决办法：

1. 将处理后的DataFrame中删掉上述两行

使用词袋countVector+LDA+机器学习方法，验证集F1-score仍然很低

（与tf-idf+机器学习方法接近，F1-score:0.34）

结论：数据处理部分存在问题”

2. 数据处理中增加对网址、单词的保留

（暂未进行）