# 公司分类任务报告

1. 任务描述

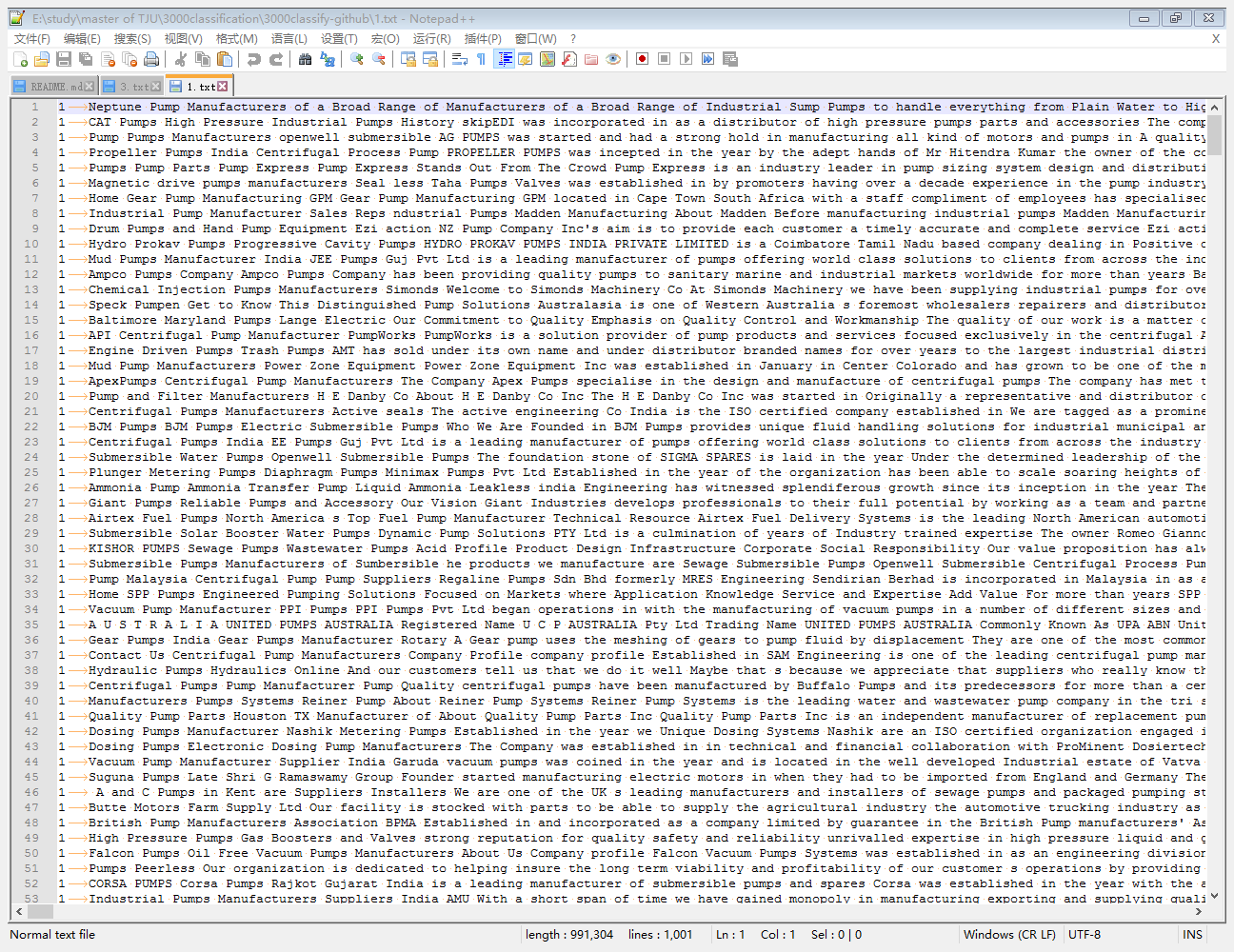
在数据集中，公司分为三类：供应链，流通链，服务链。每类公司大约有1000个样本，其中包括公司名称，公司官网网址，公司描述（有缺失值）。

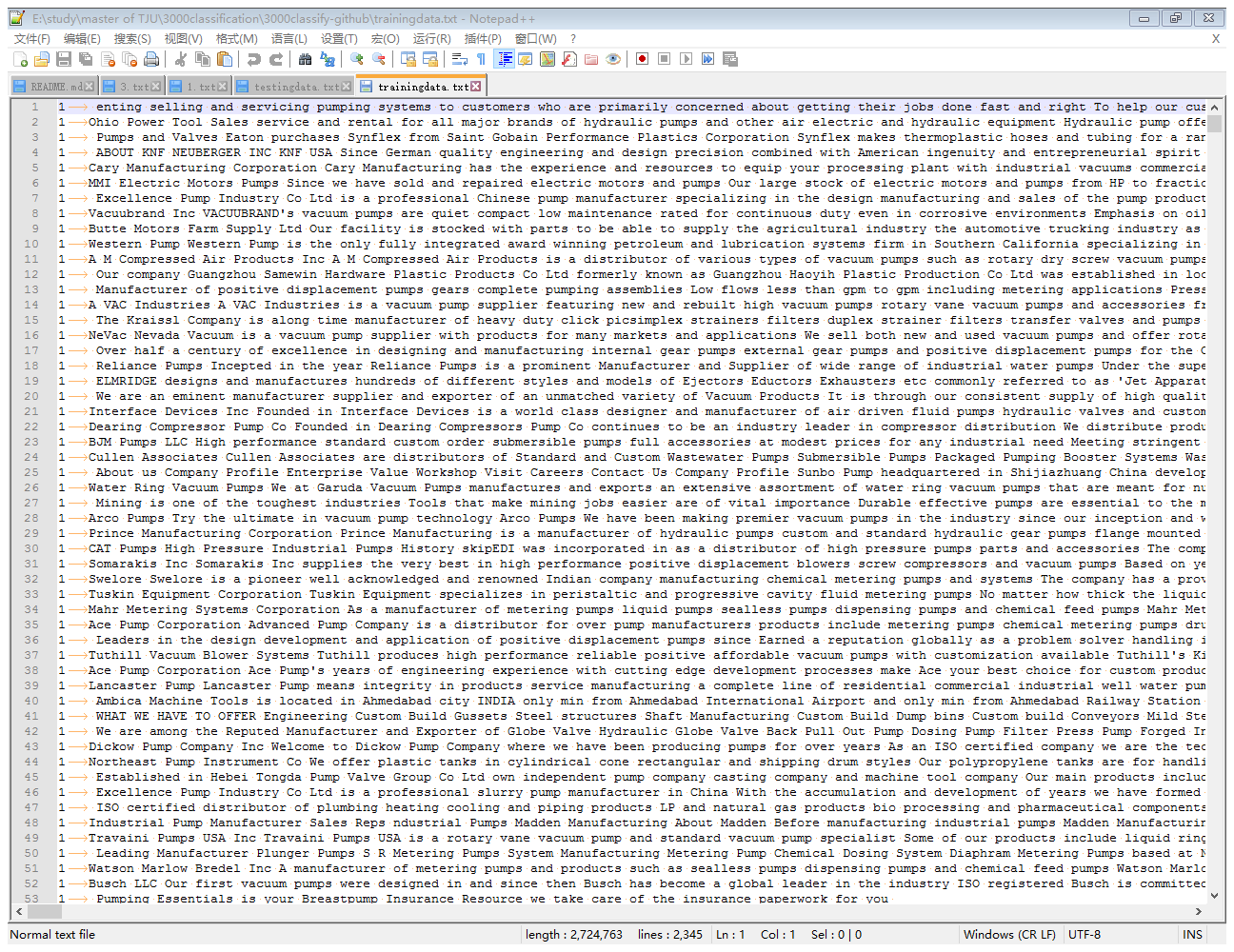
我们的任务就是根据以上信息，通过机器学习和语义分析进行分类，进行模型训练，当输入未知公司类型的公司信息时，判断其分类。

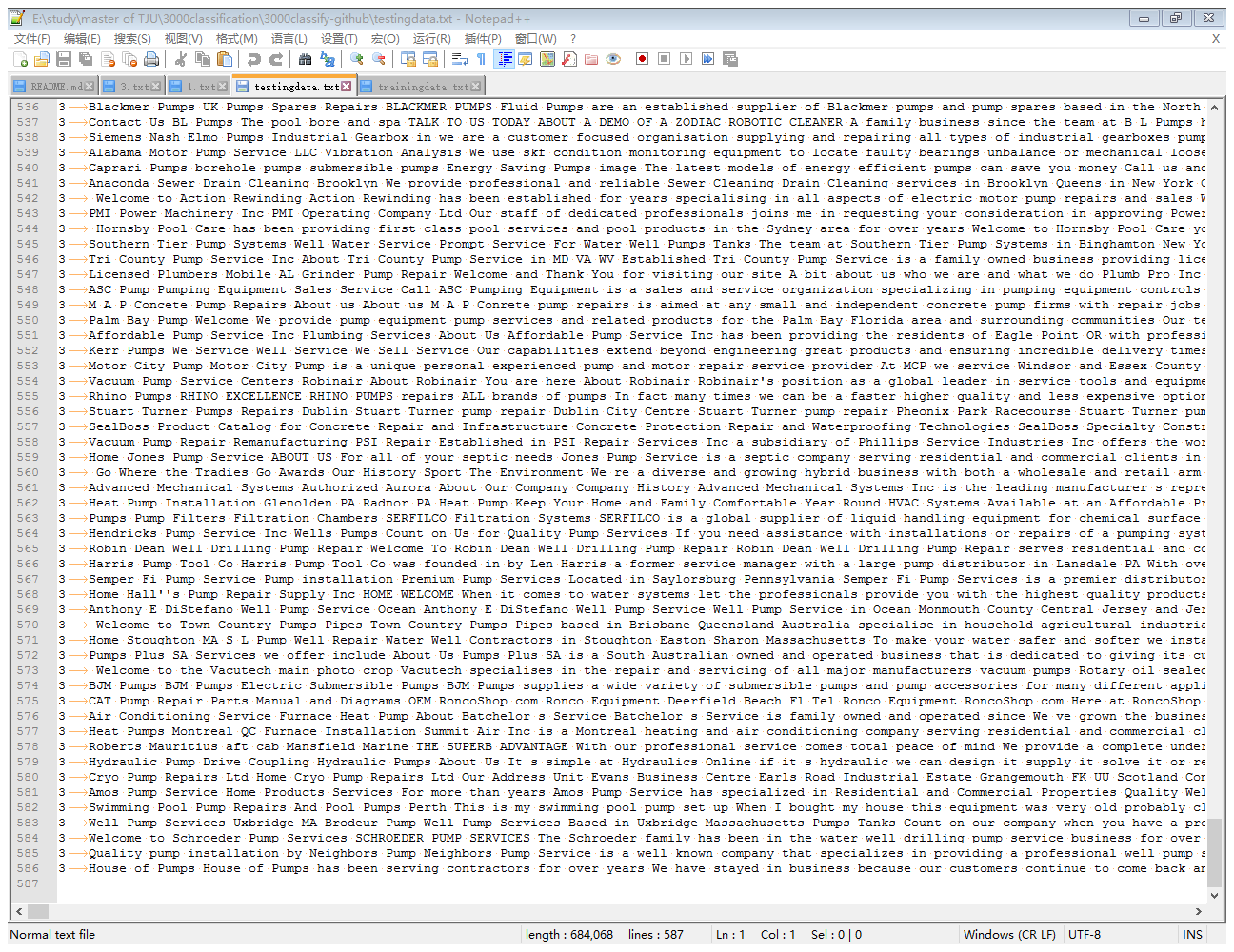
1. 数据集描述

数据集是一个Excel表格，其中包括三个表，每个表是一个类型的公司，每个表的三列分别是名称，网址，描述。

1. 基本思路及实现
2. 数据清洗：去除网址列，把三个表保存为TXT格式，把名称和描述合并成一行，前面加上标签（1 代表 供应链 2 代表 流通链 3 代表 服务链），文本中只保留字母和英文字符 ’ ，为后续预处理清除可能出现的干扰。
3. 分割数据：因为数据集不分训练集和测试机，所以需按照八二比例切割数据集，为保证随机性，我们实现自动随机分割程序，保证每类样本在训练集里有80%，在测试集里有20%（2345条训练集，586条测试集）







1. 数据预处理：去除停顿词，即英文中a、for、the等对于语义没有影响的词（约1200个），后续实验表明，去除停顿词会减少特征维度，但是有时也会影响分类效果，所以我们把这一步封装，作为可选项，在后续实验中选择。
2. 模型训练：我们选取逻辑回归模型，线性SVM（支持向量机）模型，SVM模型（线性核），SVM模型（poly核），SVM模型（sigmoid核），SVM模型（rbf核），朴素贝叶斯（NaiveBayes）模型。
3. 调整参数观察结果：我们可以选择一元模型或二元模型，是否使用TFIDF特征，是否去除停顿词，以及SVM模型中的参数。得出结果。
4. 实验结果

最一开始正确率是0.63左右，说明该数据集具有预测效应。

特征维度18311维。二元特征高达17万维。



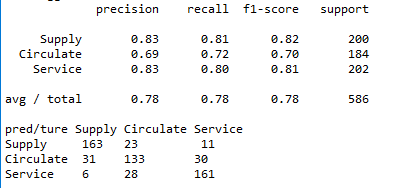
根据对比，得出最优模型，设定最优模型和参数设置：

运行100次平均准确率为： 0.780699658703

最大正确率为： 0.810580204778

下面为一次随机试验的结果混淆矩阵以便下文分析使用：

混淆矩阵：



1. 结果分析

最初各种数据预处理和参数调整都没有实现，有63%的准确率，说明在数据集有预测潜力（因为准确率明显高于33%）。

经后续调整，准确率上升至0.8左右，但特征词过多是个明显的问题。

1. 特征词过多影响分类效果：17万的特征词必然存在对分类无影响，进而冗余，造成负面影响的特征。
2. 某些公司具有两类特征：可能出现某类公司既是供应链也是服务链这种情况，所以分类并没有错，因为文本标记单一。
3. 专业词汇敏感度不够：目前实现的是一个通用性的模型，对于公司的性质我们并不了解，如果加入专业词汇词库，提取关键特征，会有帮助。
4. 不同类别表现不同：根据混淆矩阵，流通链的预测分类效果最不好，可能是统一分类模型对他的影响。
5. 后续计划

根据结果分析的原因进行改进：

1. 去除通用词，停顿词去除并不充分，根据特定背景，我们还可以去除特定通用词。
2. 需要从数据源头分析解决，判断公司是否属于两个分类或多个分类。
3. 找到特征性强的词，即出现这个词就可以直接分类。
4. 进行更精细分类，而不是三个分类，结合上一步，先进行特殊数据分类，再建立统一模型。