# 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）。其中数据经过修正得到了多标签。

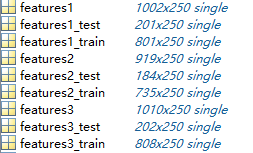
我们的任务就是根据以上信息，通过机器学习和语义分析进行多标签分类，进行模型训练，当输入未知公司类型的公司信息时，判断其分类。

1. 数据集描述

数据集是一个Excel表格，其中包括三个表，每个表是主要是某个类型的公司，每个表的三列分别是名称，网址，描述，以及是否属于三个类。

1. 基本思路及实现
2. 数据清洗：去除网址列，把三个表的文本保存为TXT格式，把名称和描述合并成一行，标签（1 代表 供应链 2 代表 流通链 3 代表 服务链）储存为三个.mat文件，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. 词向量训练：我们采取词向量表示方式，进行语义特征提取。例如，I am in the house 和 I am in the restaurant， 其中由于house和restaurant因为在句子中的的位置相似，且前面的词一直，所以这两个词是相近的词，的空间向量相似程度高。最后得到一个表，每个词都由一个250维的向量表示。
4. 句子向量训练：在词向量的基础上，根据句子中的词，把句子转换为向量表示形式，也是250维表示一个句子，作为每个公司的特征。
5. 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%（2344条训练集，587条测试集）
6. 模型训练：我们选取朴素贝叶斯（Naïve Bayes）多分类模型。
7. 调整参数观察结果：模型训练中的PCA降维比例，词向量训练中的维度和窗参数等参数对最终结果都有重要影响。
8. 实验结果

特征维度250维，窗参数4维，PCA降维比例为10%，



根据对比，得出最优模型，设定最优模型和参数设置：

运行150次平均准确率为： 0.807962784805970

最大正确率为： 0.84（作为示例程序，训练集和测试集的选取已经储存）

下面为最后一次试验和平均实验的结果：



1. 结果分析

是文本特征上升到语义特征后，准确率上升至0.81左右。

拿到新数据后，我们首先尝试了构建特征词库和句向量的方法进行二分类，但效果并不是很好，说明二分类会出现误区，所以我们构建了多分类方法。

在后续多分类的问题上，参数选择是个问题，对结果有重要影响。

1. 向量维度影响分类效果：选择多少维的向量表示对于一句话的语义表示和之后的分类效果有不同程度上的影响。
2. 窗参数的选择：如何得到一个词的向量表示，这就是窗参数的作用，使用前后的几个词计算一个词的向量对整体性能表现很重要。
3. 不同类别工作表现不同