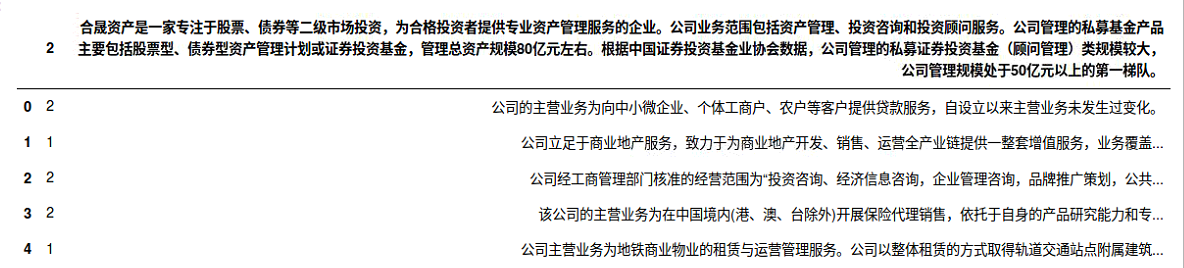
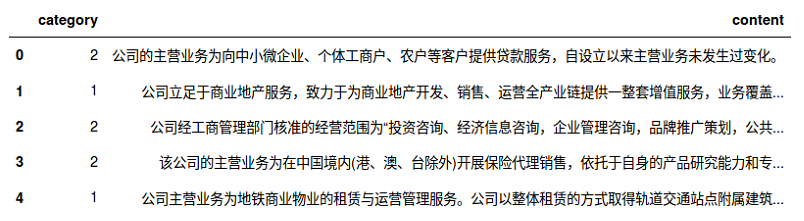
**第一次汇报**

程序见：my\_project\_text\_classification\_4

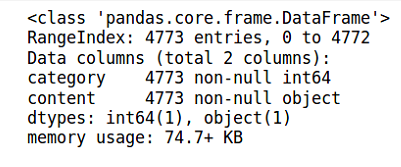
1. 数据探索
2. 数据分布与分析



将其列名修改为category、content：



此数据集的整体大致内容为：



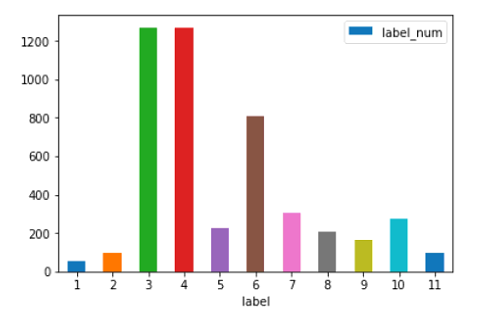
Category:

[1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11]

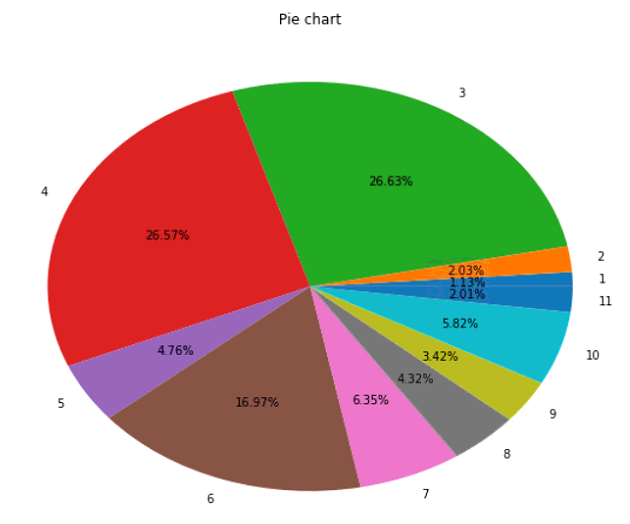
Category\_num:

[54, 97, 1271, 1268, 227, 810, 303, 206, 163, 278, 96]

条形图：



百分比情况：



二、采用jieba进行分词

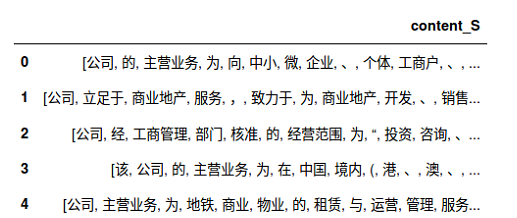
加载了现代汉语词库

调整字典每个词语词频

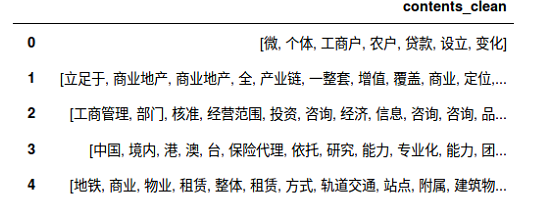
分词

加停用词表后再分词（词频>1000）

未加停用词分词的效果:



加停用词分词的效果:

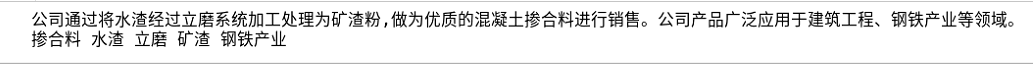


1. 分词后的数据展示
2. 词云

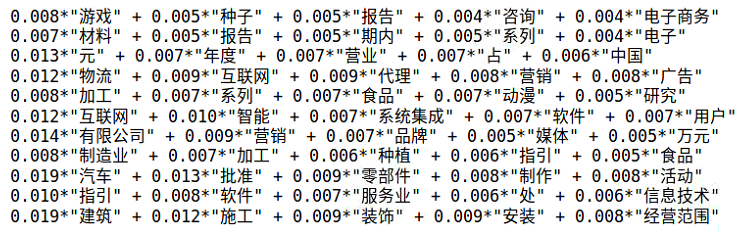


1. TF-IDF：提取关键词

第2411条数据:



1. LDA：主题模型，查看11个分类大致代表什么含义



1. 词向量和模型

根据词频做词向量(countVector)、TF-IDF做词向量(10折交叉验证)

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 方式 | KNN | 贝叶斯 | SVM | 随机森林 | 神经网络 |
| 词频(CV) | 0.57 | 0.843 | 0.814 | 0.536 | 0.8315 |
| TF-IDF | 0.819 | 0.768 | 0.85 | 0.766 | 0.84 |

1. 计划

|  |  |
| --- | --- |
| 第二周 | CNN |
| 第三周 | 模型融合 |
| 第四周 | 总结 |