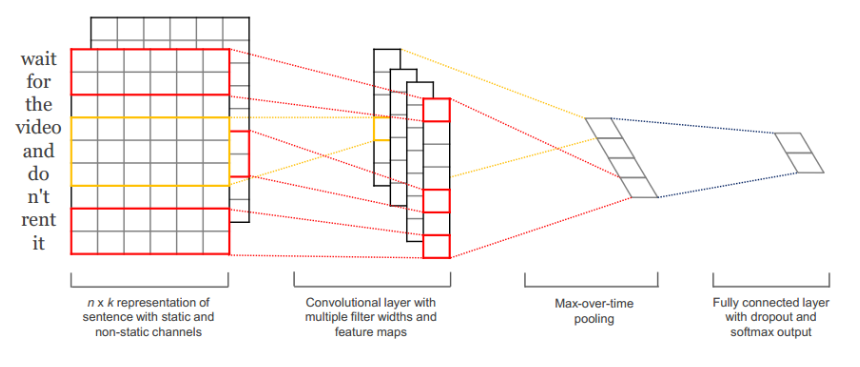
**第三次汇报**

**testing程序见：****my\_project\_text\_classifcation\_test\_7、cnn程序在text\_classification文件下**

1. **CNN  模型**



Sentences = 4773, 样本数。

max\_document\_length=570，选取每个样本中分词后的最大词个数，不足就采用补齐的方式。

filter\_sizes = “3,4,5”

num\_filters = 128.

Pool\_size = max\_document\_length- filter\_sizes+1

采用洗牌的方式

训练集（90%）：准确率，95%

验证集（10%）：准确率，75%

1. **选取最终模型**

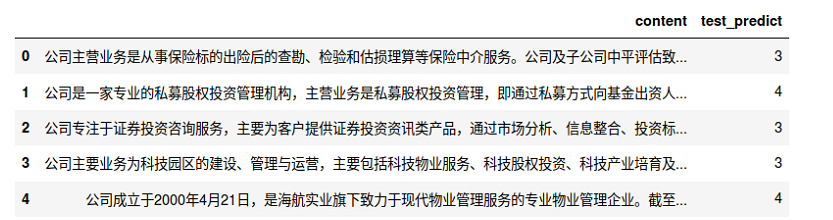
|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 方式 | KNN | 贝叶斯 | SVM | 随机森林 | 神经网络 |
| 词频(CV) | 0.566 | 0.8564 | 0.8120 | 0.8080 | 0.8470 |
| TF-IDF | 0.8214 | 0.7628 | 0.8522 | 0.8091 | 0.8602 |

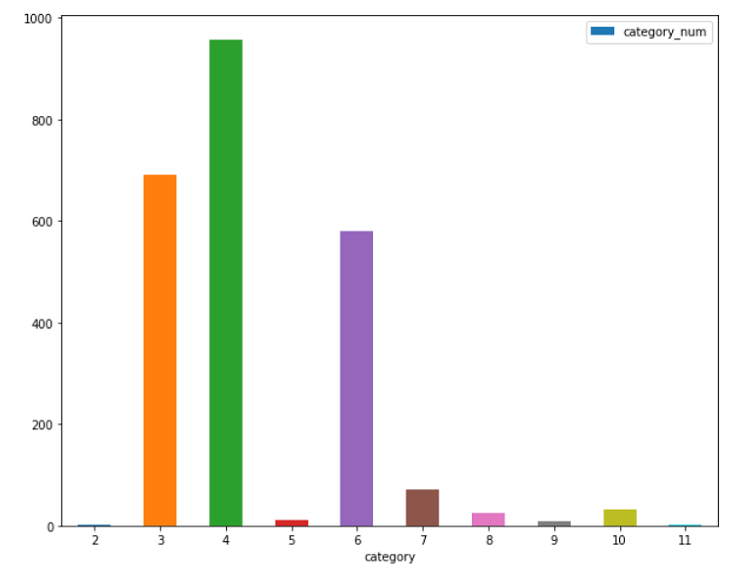
CNN: 75%

最终选取：神经网络

1. **预测结果**

**相关代码放在：my\_project\_stopwords\_test\_7.ipynb**





1. **总结：**
2. 通过这次项目知道了自然语言处理文本分类开发的流程。
3. 词向量的构造（词频、TF-IDF、Word2vec）
4. 知道了常用的机器学习算法在NLP上的应用（KNN,SVM,Randomforest, Bayes, neural\_network,CNN）
5. 知道了自己在开发方面的不足（比如：代码的耦合性不是很高）