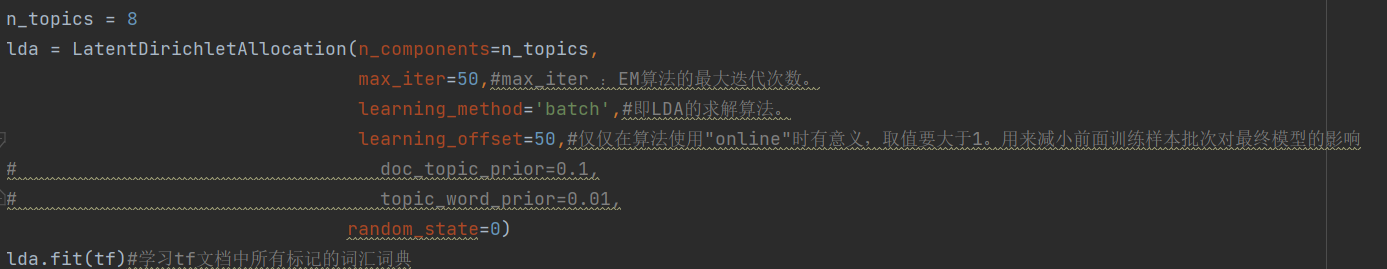
本周进行的工作内容

1.在上周的基础上对获取到的文本数据进行了进一步的词语分割和词性标注。

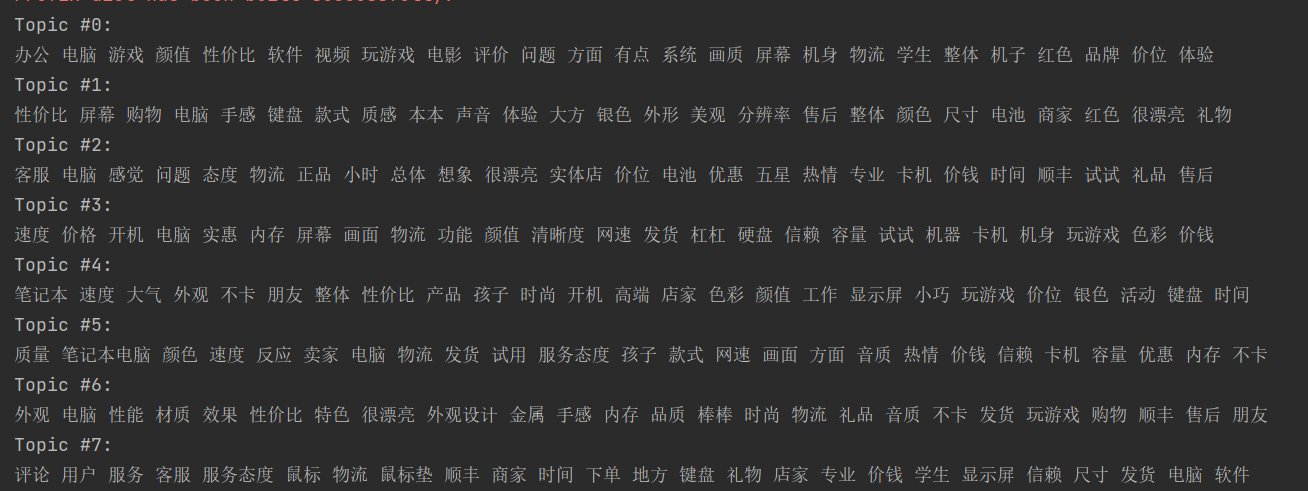
2.学习了Python-sklearn中的CountVectorizer类。知道了他是一种常见的特征数值计算类，是一个文本特征提取方法。对于每一个训练文本，它只考虑每种词汇在该训练文本中出现的频率。还了解到了CountVectorizer会将文本中的词语转换为词频[矩阵](https://so.csdn.net/so/search?q=%E7%9F%A9%E9%98%B5&spm=1001.2101.3001.7020" \t "https://blog.csdn.net/weixin_38278334/article/details/_blank)，它通过fit\_transform函数计算各个词语出现的次数。

3.继续学习了LDA模型相关的基础知识和算法思想，得知了Python中的LDA(LatentDirichletAllocation)函数的相关参数意义，并以自己的语言进行了标注。





4.利用LDA模型算法,初步遴选出了一定数量的分好类的文本主题小组



5.将上述生成的文本主题数据进行了本地保存。

