7.5 实战练习

在其余参数全部固定不变的情况下，尝试分别用清理前矩阵、清理后原始矩阵、TF-IDF矩阵进行LDA模型拟合，比较分析结果。

在gensim拟合LDA时，分别将passes参数设置为1、5、10、50、100等，观察结果变化的情况，思考如何对该参数做最优设定。

请尝试对模型进行优化，得到对本案例较好的分析结果。

> 提示：使用gensim进行拟合更容易一些。

8 文档相似度

8.1 基本概念

8.2 词条相似度：word2vec

词袋模型不考虑词条之间的相关性，因此无法用于计算词条相似度。

分布式表达会考虑词条的上下文关联，因此能够提取出词条上下文中的相关性信息，而词条之间的相似度就可以直接利用此类信息加以计算。

目前主要使用gensim实现相应的算法。