一．摘要

LDA（Latent Dirichlet Allocation，隐含狄利克雷分布）主题模型的主要功能在于预测文档的的主题分布，是一种非监督机器学习技术。该方法将一个文档集分为文档、主题和词三个层次，LDA 模型的主要步骤是:

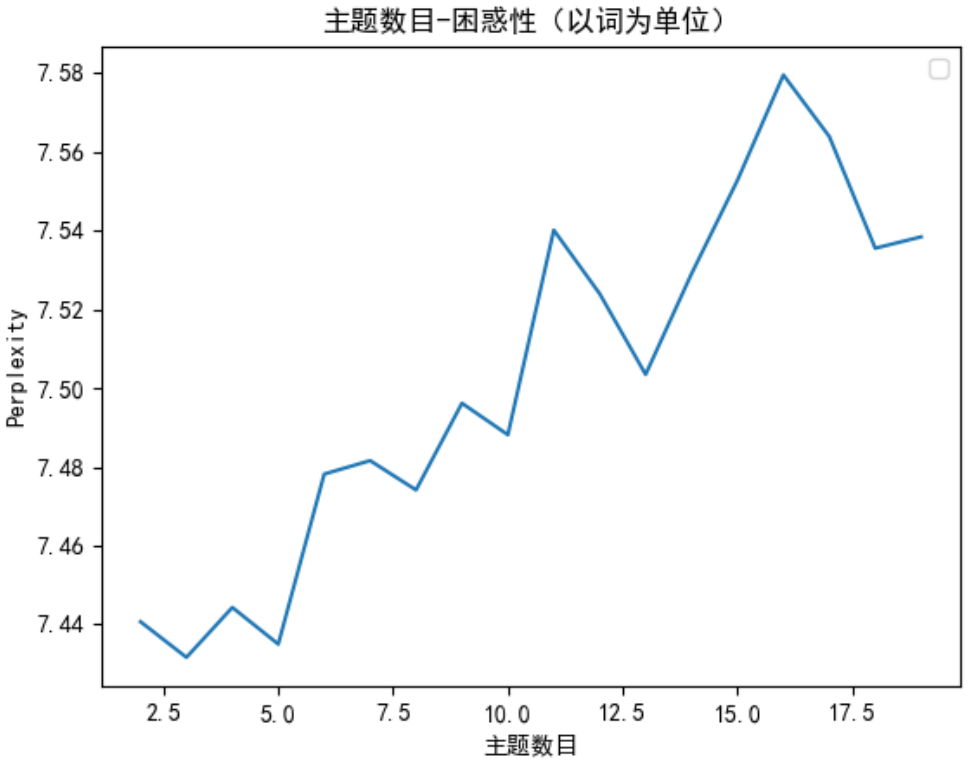
1. 确定主题数目，根据狄利克雷分布确定每个主题和词汇的分布;
2. 根据狄利克雷分布确定每个文档和主题的分布;
3. 遍历每个文档中的每个单词，根据上面两个分布（第一步的分布是由先验知识得到的），重新分配其所属的主题，改变调整两个分布，不断迭代直到模型收敛。

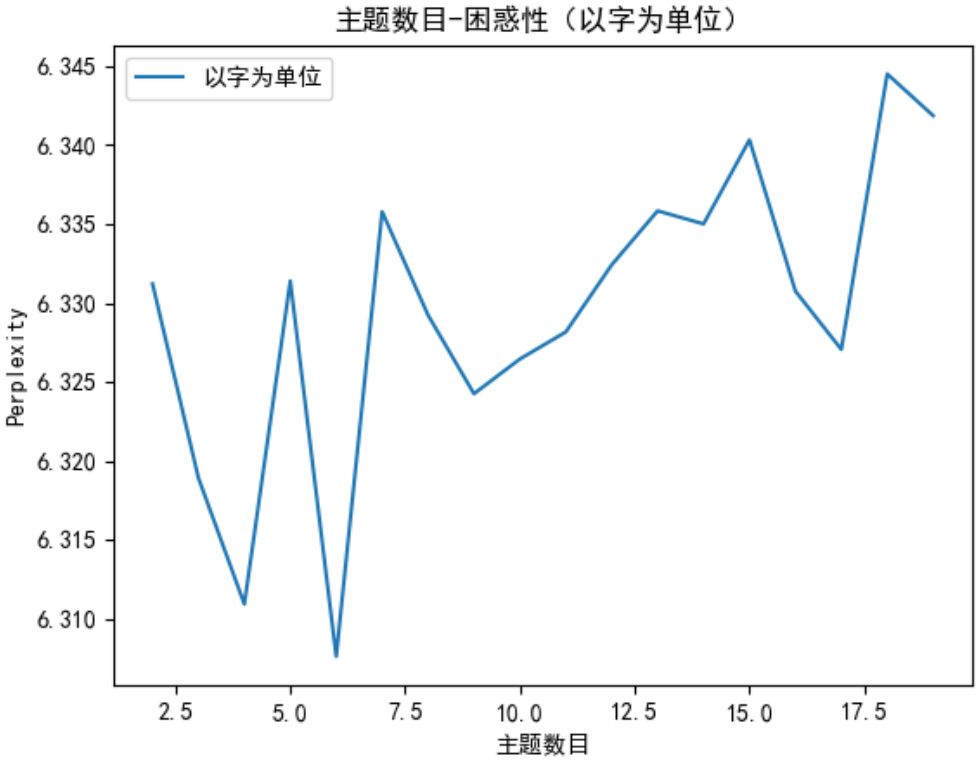
LDA 模型的质量可以使用指标困惑度（perplexity）和主题一致性（coherence）来衡量，困惑性越低或者一致性越高说明模型越好，也可以使用此方法来确定最佳的主题数目。值得注意的是，困惑性存在可比性问题，不同分词方式得到的 LDA 模型间的困惑性不应被用于比较优劣。

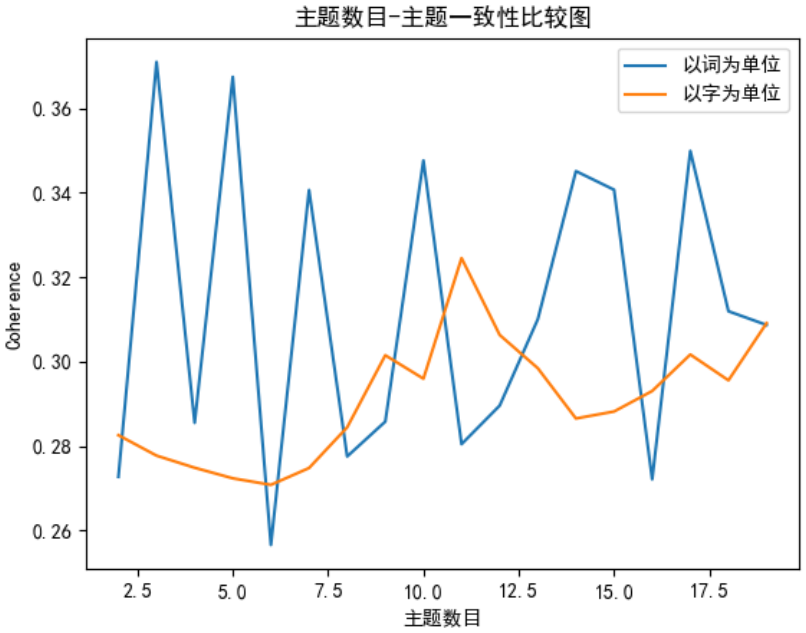
二．实验过程及结果

选择使用 Python 为编程语言，使用 jieba 库进行分词工作。由于最初的文本中含有标点符号和网址广告等不相关信息，需要首先进行数据预处理工作。经查阅资料，使用 Gensim 库这一简单高效的自然语言处理 Python 库。由于个人电脑性能受限，主题数设置为 2 到 20 之间的 19 个整数。

下面展示不同分词方式和不同主题数目下的 LDA 模型困惑性和主题一致性结果:







三．结论

不论是以词还是以字为单位进行划分，LDA 模型均在 15-18 范围内取到最小的困惑性（程序中对困惑性数值进行了取反），这和文档集中实际的主题数目 16 是比较吻合的。

同时，通过分析两种分词方式的主题一致性，可以看到，以词为单位的划分方式总体要优于以字为单位的划分方式，但波动较大，这可能是因为以字为单位划分会有更多与主题关联不大的重复，导致 LDA 模型质量变差但却更稳定。