**基于LDA模型的主题建模**

郭润堂

guoruntang@buaa.edu.cn

**Abstract**

LDA（Latent Dirichlet Allocation，隐含狄利克雷分布）模型为一种主题模型，可以用概率分布的方式表示文档集中每篇文档的主题。本文采用了 LDA 模型对不同小说的段落进行了主题建模和分类，从 16 本中文小说中均匀抽取了 200 个段落作为数据集，每个段落的长度大于 500 个字，且标记了其所属小说的类别。在不同主题数下对段落进行了分类，并评估了分类性能。实验结果显示，选择合适的主题数对提高分类准确率有重要作用，本文选取了一致性最高的主题数12进行了进一步分析。再次随机抽取了20个类，对模型分类的准确性进行评估。实验发现以"词"为基本单元的方法优于以"字"为基本单元的方法。

**Introduction**

LDA模型是一种概率主题模型，由Blei, Ng 和 Jordan 于2003年提出，用来发现文档集合中的隐含主题。LDA模型的基本假设是，文档是由多个主题混合而成的，每个主题又是由多个词语构成的。LDA模型的目标是，根据文档集合，推断出每个文档的主题分布和每个主题的词语分布。LDA模型包含文档集合、词典、主题词语分布、文档主题分布等四个基本元素。文档集合包含多篇文档，每篇文档由多个词语组成。词典：所有文档中出现的不同词语的集合。主题词语分布：包含多个主题，每个主题是一个包含多个词语的多项式分布，每个词语在主题中的权重表示该词语与该主题的相关程度。文档主题分布：每篇文档由多个主题混合构成的多项式分布，每个主题在文档中的权重表示该主题在文档中的重要程度。

LDA模型是一个三层贝叶斯概率模型，包含词、主题和文档三层结构。其生成过程如下：

1）对于每个主题，从一个狄利克雷分布中采样出一个词语分布；

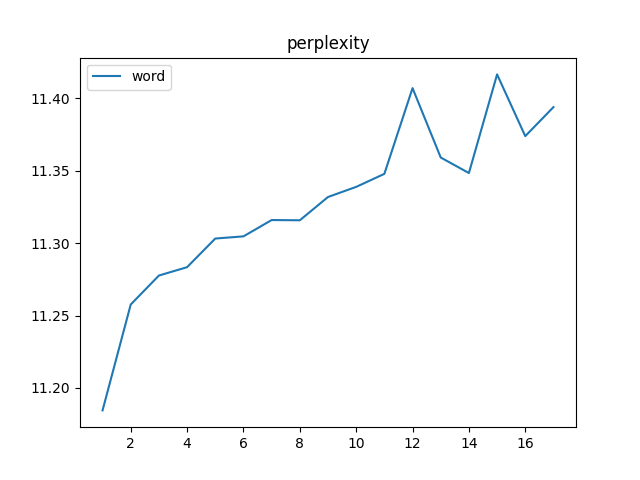
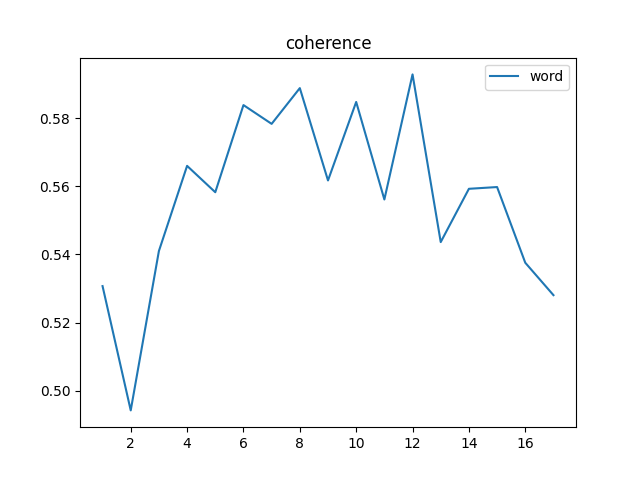
2）对于每篇文档，从一个狄利克雷分布中采样出一个主题分布；

3）对于文档中的每个词，先从主题分布中采样出一个主题，再从该主题的词语分布中采样出一个词。

LDA模型的训练过程是一个迭代的过程，需要用到贝叶斯推断和马尔可夫链蒙特卡罗方法。常用的方法有EM算法和吉布斯采样等。LDA模型的输出结果包括每个主题的关键词语，每篇文档的主题分布以及每个词语的主题分布。可以通过这些结果来进行话题分析、文本分类、推荐系统等应。

**Methodology**

测试中首先均匀抽取200个字数大于500的段落，进行去除停词处理。在主题数目由1至18测试复杂度与一致性结果如图：

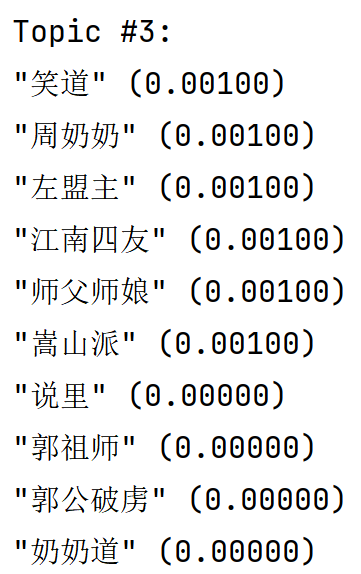
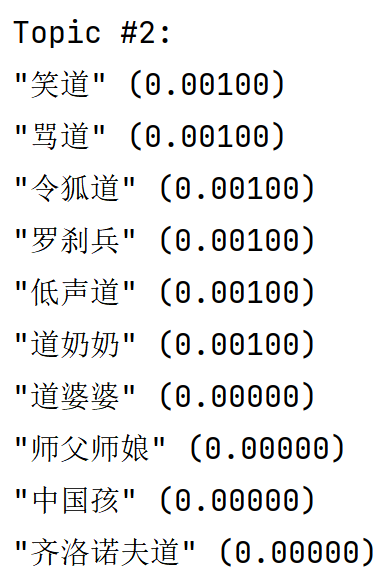
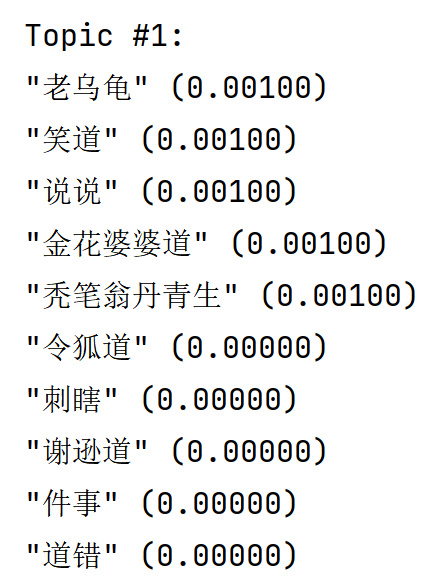
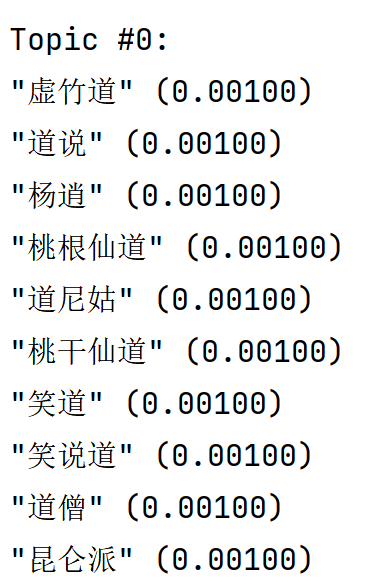
数据如下：

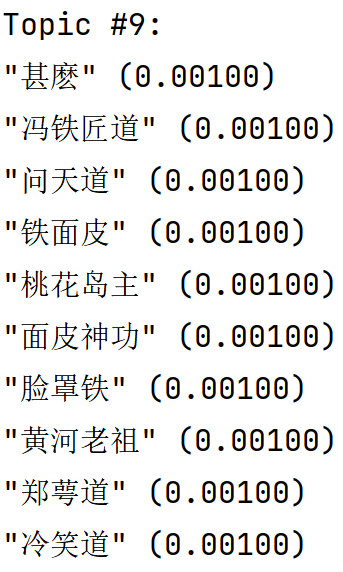
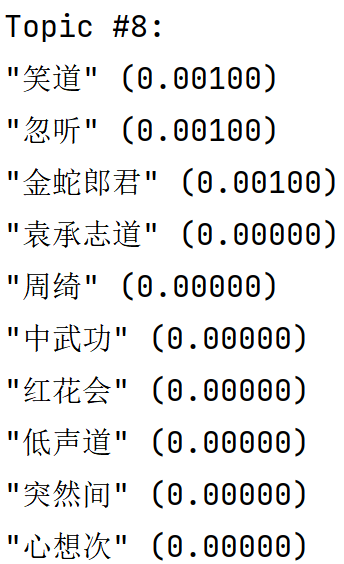
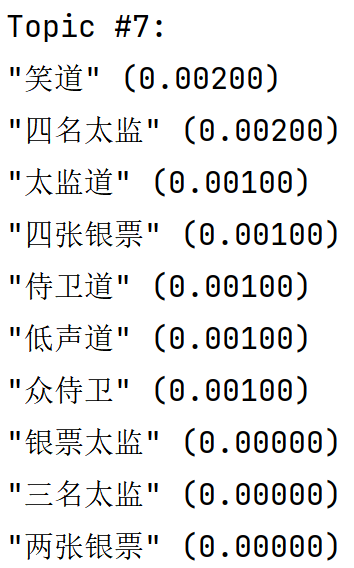
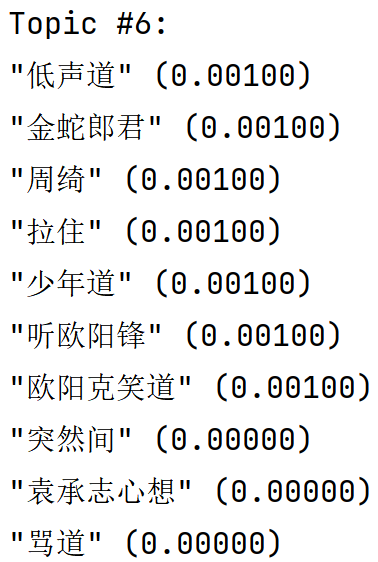
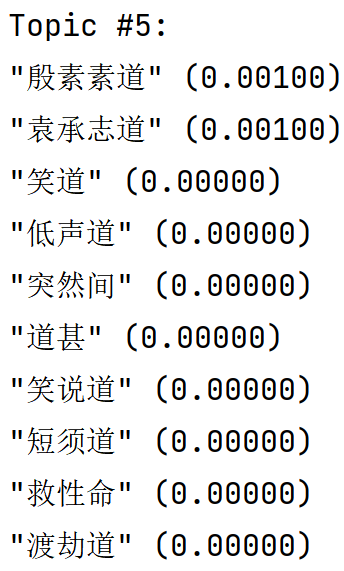
results\_perplexity\_word : [11.184519060435806, 11.257669014427202, 11.277671712716419, 11.283446238263581, 11.30322977331532, 11.304730659162487, 11.316000501449318, 11.31583806991786, 11.331920244276375, 11.33891425584064, 11.3479055577709, 11.40713455866992, 11.359199920978684, 11.348474461753774, 11.416604672613891, 11.373971986129614, 11.394003081025641]

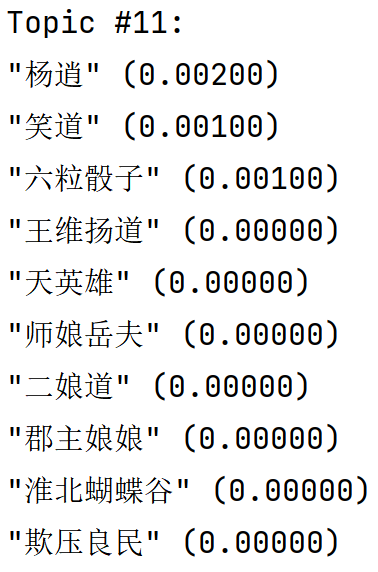
results\_coherence\_word : [0.530704842940443, 0.494156927838228, 0.54102901774359, 0.5660573740223206, 0.5583051465344642, 0.5839290408058517, 0.5784055306879463, 0.5889188002720958, 0.5617666675536006, 0.584847601948411, 0.5561715605688281, 0.5929486220037135, 0.5436454863361445, 0.5593107098526457, 0.5598282056762185, 0.5375700105439212, 0.5280425413065456]

由于第12类的一致性最高，在分类为12类的基础上进行进一步测试：

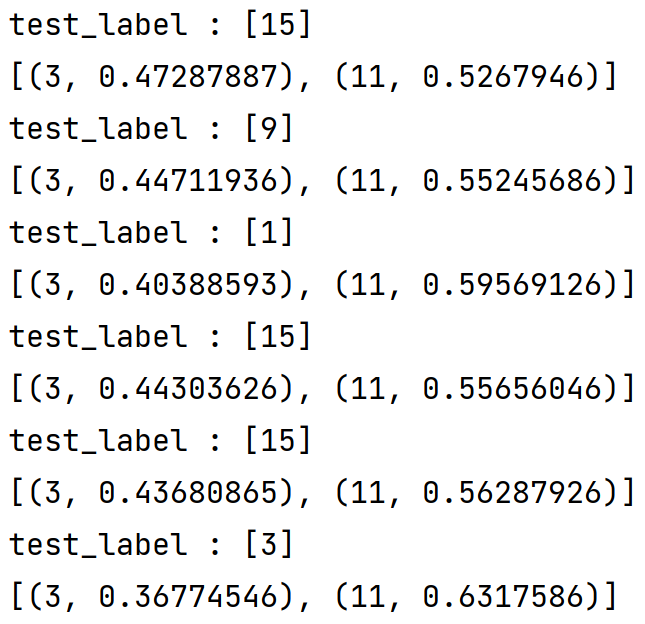
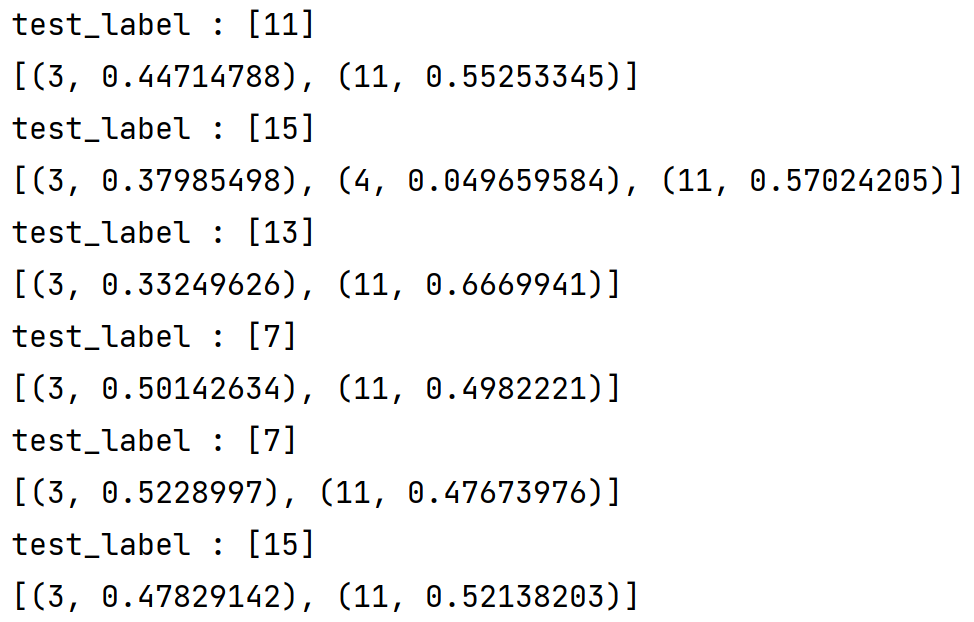
分得的12类，最高频的10词分别为：

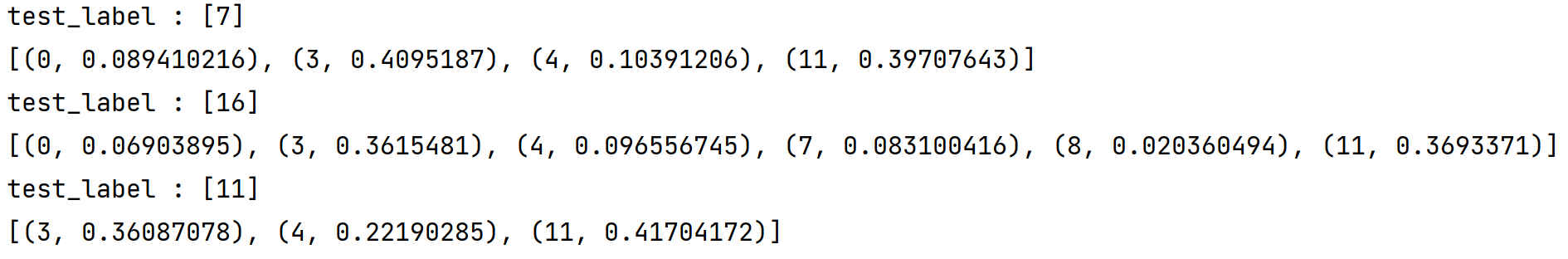


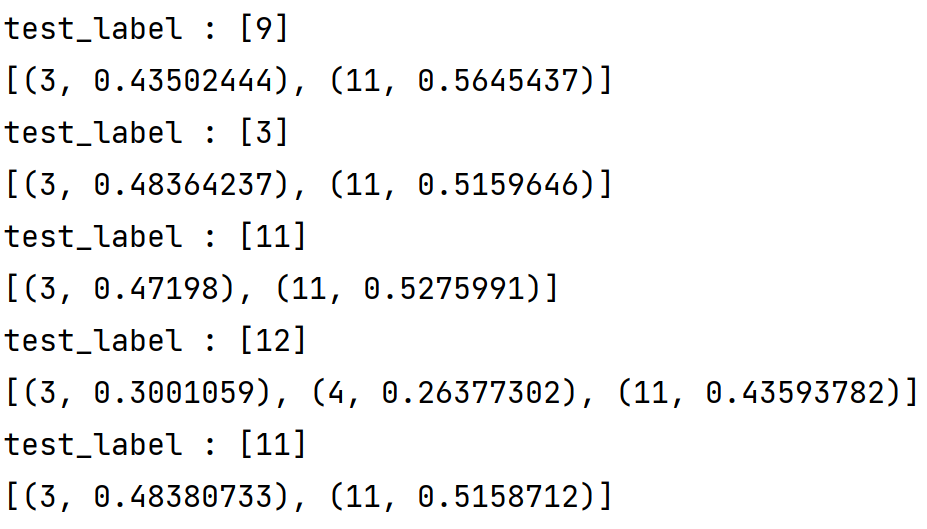




对随机抽取的20段超过500字的文本进行相同处理后进行测试，test\_label为原标签，下为预测类别及概率：







**Conclusions**

本文通过 LDA 模型对中文文本进行主题建模，并探究了不同的基本词元和主题数对于困惑度、主题一致性的影响。实验结果表明，在使用 LDA 模型对于中文小说进行主题建模时，选择“字”作为基本单元相对于“词”作为基本单元可以得到更小的困惑度。同时，当主题数较少时，分类准确率相对较高，但过多的主题数会导致模型过拟合，准确率反而下降。