NLP_LDA大作业

ZY2103203 黄旭聪 github:https://github.com/189223/NLP_Assignment3.git

1. 问题阐述

从给定的语料库中均匀抽取200个段落（每个段落大于500个词）， 每个段落的标签就是对应段落所属的小说。利用LDA模型对于文本建模，并把每个段落表示为主题分布后进行分类。验证与分析分类结果。

LDA指代Latent Dirichlet Allocation, 是由Blei, David M.，Andrew Y.，Jordan于2003年提出的。其目的是用来推测文档的主题分布。它可以将文档集内的每篇文档的主题以概率分布的形式给出，从而通过分析文档及抽取主题分布后，根据主题分布实现主题聚类或文本分类的任务。**这里应用金庸小说集实现文本分类的任务。**

2.背景知识

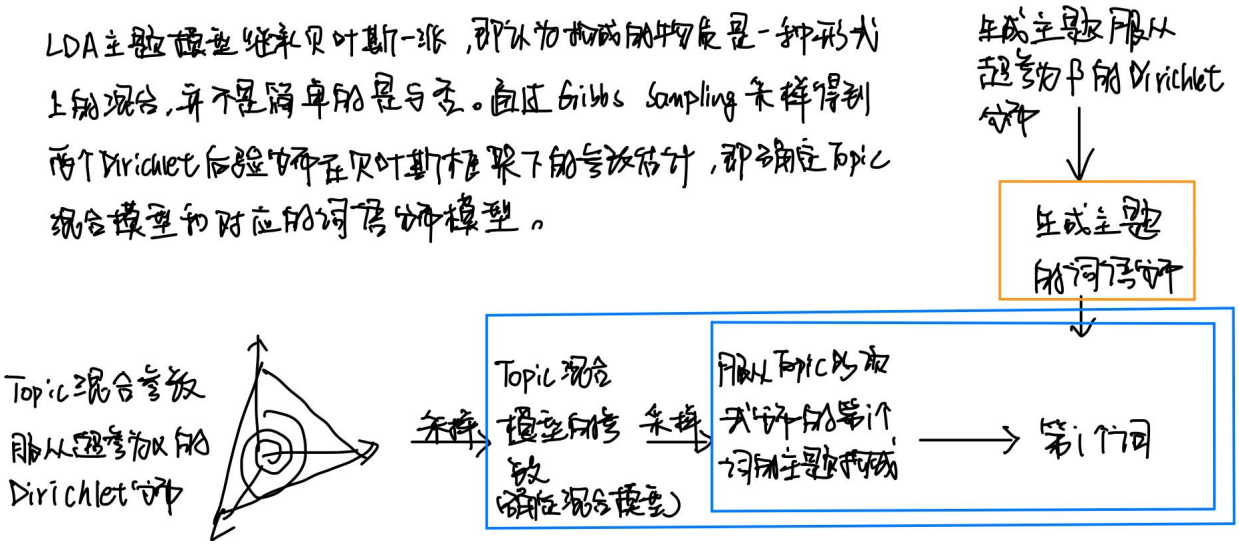
- **词袋模型**

LDA 采用词袋模型。所谓词袋模型，是将一篇文档，我们仅考虑一个词汇是否出现，而不考虑其出现的顺序。在词袋模型中，“我喜欢你”和“你喜欢我”是等价的。与词袋模型相反的一个模型是n-gram，n-gram考虑了词汇出现的先后顺序。

- **LDA文档生成过程**

1. 按照先验概率选择一篇文档
2. 从Dirichlet分布中生成文档的主题分布，主题分布由超参数为α的Dirichlet分布生成
3. 从主题的多项式分布中采样生成文档第i个词的主题
4. 从服从另一个超参β的Dirichlet分布中采样生成第i个词的主题对应的词语分布
5. 从词语的多项式分布中采样生成第i个词

更为具体的LDA文档生成过程如下手推过程：



- **基于LDA主题模型的文本分类**

1. 数据集生成

   题目要求产生200个段落，用于生成LDA模型。然而金庸小说数据集中，有16本小说，无法均分的产生样本。为了减小先验信息的干扰，选用了文本量排名前十的小说作为数据集。每个段落包括数据清洗后的有效500词。总共随机截取10部小说的各20个段落作为训练集；随机截取10部小说的各2个段落作为训练集，即训练样本数：测试样本数=4：1，符合验证评估的要求。

2. 训练LDA模型

   用训练集训练并保存LDA模型，其中涉及到超参-主题数的确定。因此，总共生成主题数从5到60变化的LDA模型，通过以下评价指标进行超参的选择。

   困惑度（perplexity），指的是在文本分析中，训练出来的模型识别某些文档包含哪些主题具有不确定性，是不确定性的度量。 因此数值越低，不确定性就越小，则最后的聚类结果就越好。

   主题一致性（coherence），主要是用来衡量一个主题内的词是否是连贯的，即是否这些词是相互支撑的。

3. 文本分类

   - 基于LDA模型，根据确定的主题数，段落可以转变成这些主题的描述，即主题分布。提取主题分布信息作为描述段落的相关信息，以段落来源作为标签，实现文本分类任务。

- 以提取的信息构成训练样本，运用SVM进行分类。考虑到线性不可分的问题，应用非线性模型，模型性能评价指标为分类准确度。

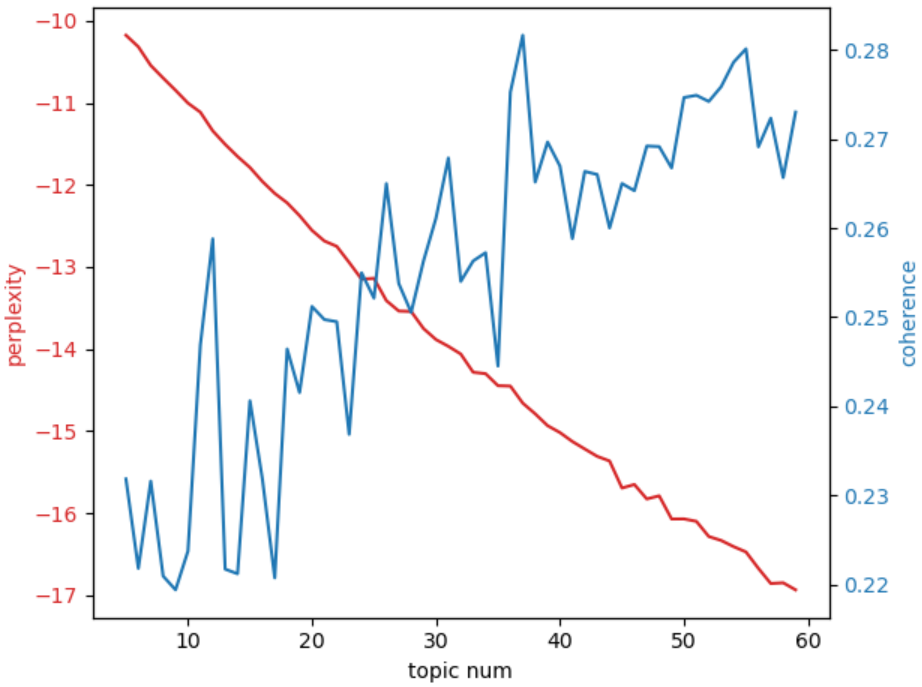
4. 预测

   在测试集上完成文本分类预测。通过训练好的LDA模型，提取测试集的主题分布；送入训练好的SVM网络，得到预测的文本类别。该指标作为性能评价的最终指标。

3.实验结果

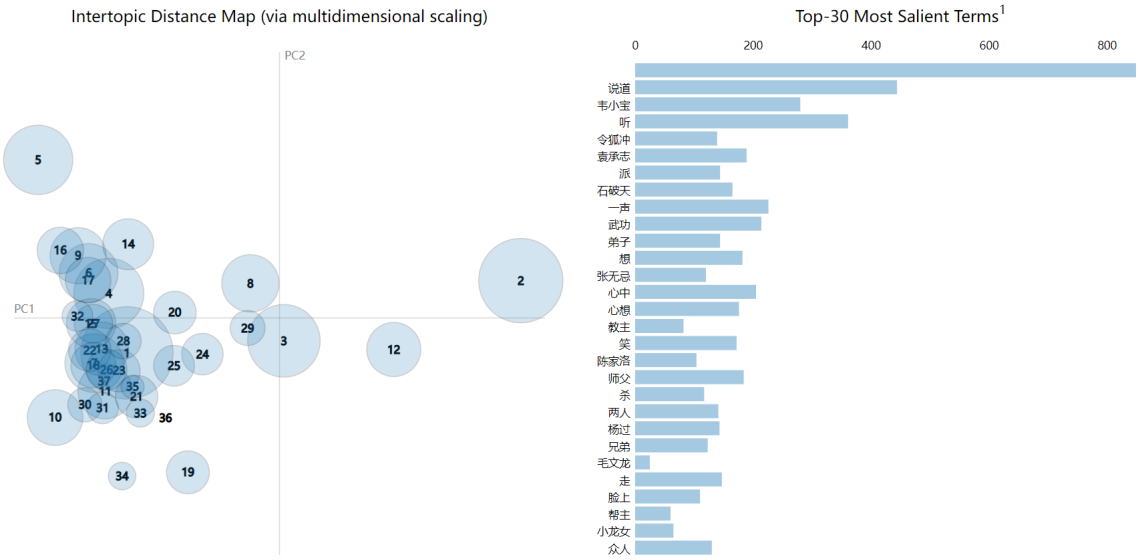
- 主题数确定

生成主题数从5到60变化LDA模型对应的困惑度及一致性指标：

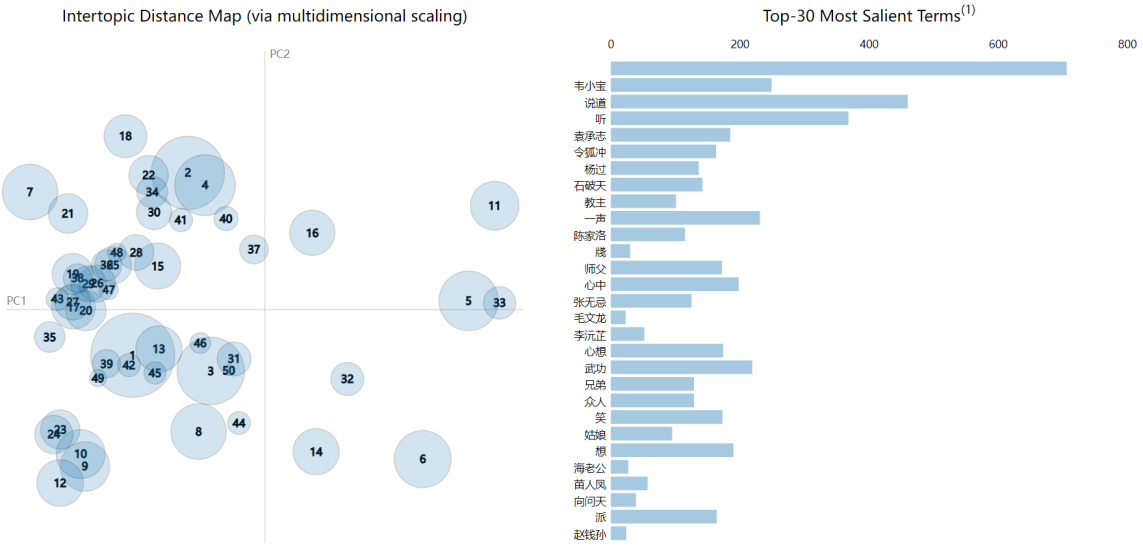


通常，超参的确定通过困惑度判断即可。但是可以发现，随着主题数的增多，困惑度越来越低。但是这并不意味着选择主题越多越好。因为当主题太多时，模型存在过拟合的问题。因此，需要通过一致性指标进一步确定。

为了进一步验证选择主题数的合理性，选取一致性最高的主题数37，56，进行更为直观的可视化显示：



主题数为37的分类结果



主题数为56的分类结果

左图是所有主题进行降维后，前两个主成分构成的映射图。每个圆形对应一个主题，半径大小对应着相应主题的比重。右图为主要的词语。相对比可以发现，主题数为56主题分离的更开，相关性更低；并且前30个高频词语中较少的存在无意义的常用词语，例如（走，两人等）；而是能识别出更多的专有名词，例如（苗人凤、向问天等）。

- 模型评估

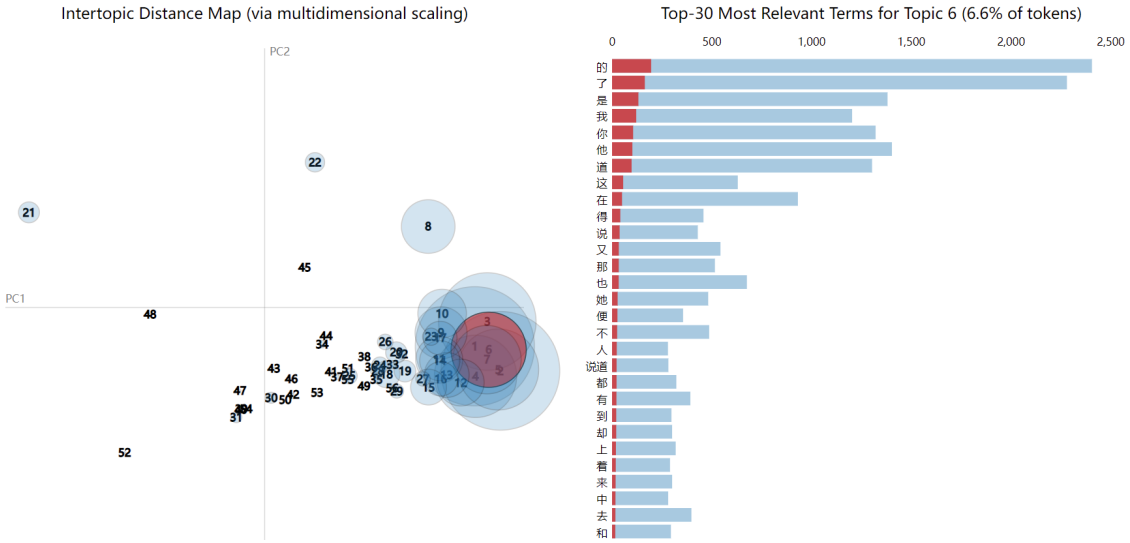
关于选用的主题数，是否去除停用词进行了一系列的对比，相关结果如下表：

主题数	单个段落包含字数	是否去除停用词	训练集分类准确度	测试集分类准确度

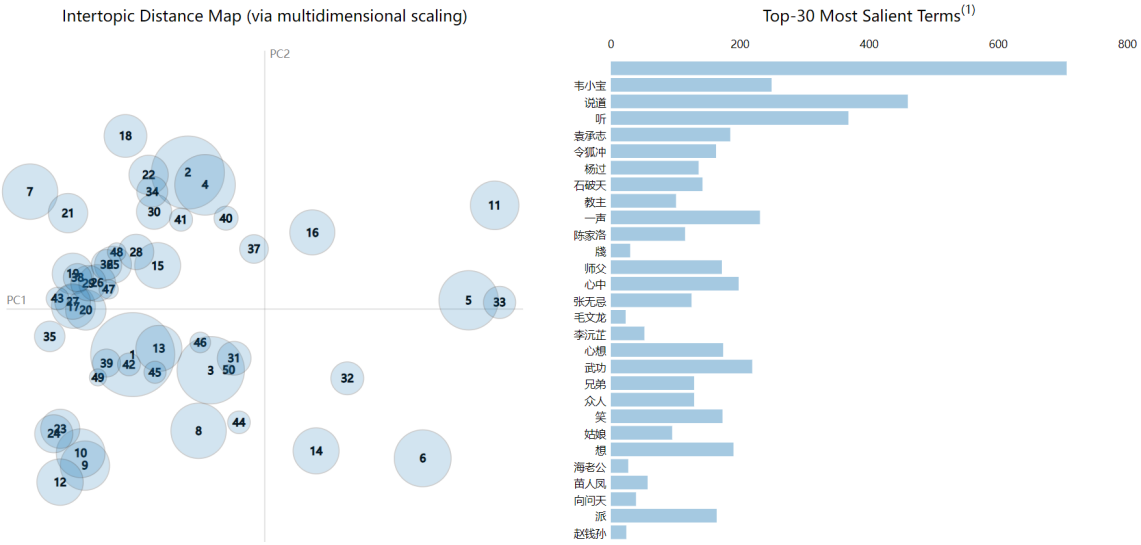
37	500	否	29%	2.1%
37	500	是	46%	15%
56	500	否	37%	20%
56	500	是	53%	40%

- 针对性优化

首先为了对比说明常用词语的影响，画出了没有去除停用词和去除停用词后的可视化图形：



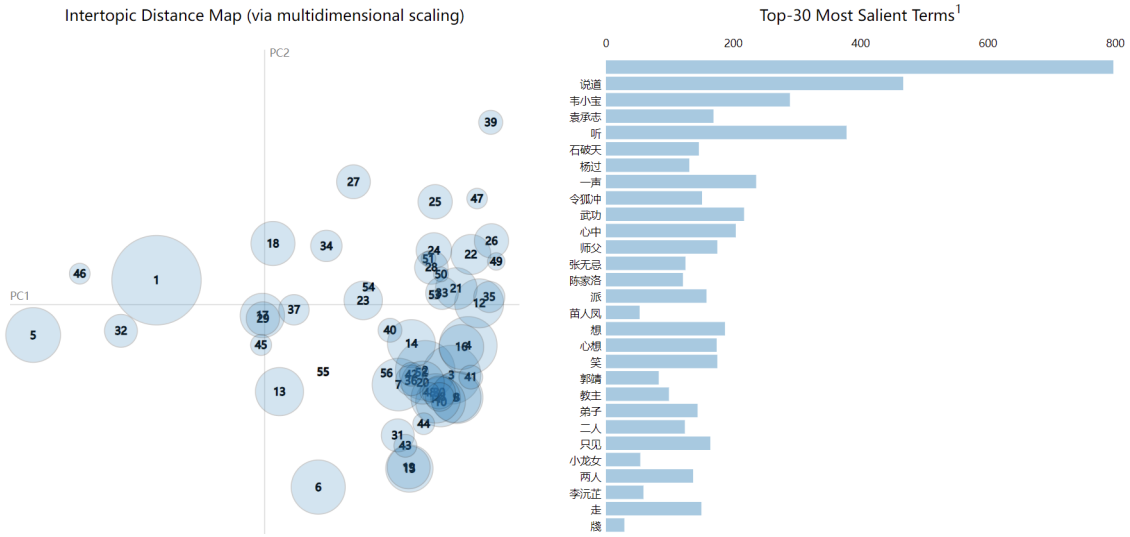
主题数56，未去除停用词的主题模型



主题数为56，去除停用词的主题模型

对比去除停用词与否的原因并不是想说明去除停用词的作用，而是为了说明常见词语对于训练主题模型的影响。许多常见词语出现频率很高，往往会掩盖了有用信息。

因此，考虑到TF-IDF模型考虑了词频，所以将词袋改为了基于TF-IDF的模型，主题模型结果如图



主题数为56，引入TF-IDF的主题模型

相关的预测结果如下表：

主题数	单个段落包含字数	是否引入TF-IDF	训练集分类准确度	测试集分类准确度
56	500	否	53%	40%
56	500	是	54%	45%

4.讨论

- 由模型评估的相关对比实验说明：1.主题数对于文本分类准确度影响很大。主题数的选择评估流程很有必要，如果随机选择，其结果很难得到较优解。2.停用词的选择很有必要，当前的停用词表是通用的表格，如果针对金庸小说加入专门的停用词，预测准确性必将更高。3.本次实验没有对比段落数、段落长度的影响，但是容易得，随着段落数、段落长度的增加，预测准确性势必会提升。
- 由针对性优化对比实验可以发现，词袋模型没有考虑常用词的词频影响，更有价值的专门名词容易滤掉，例如郭靖等主人公名称。课程中提到了TF-IDF模型，该模型考虑了词频的影响，故在LDA中考虑了词频的影响，可以发现，主题的分离程度变好，训练集分类准确度有改善，测试集分类准确度有较大改善。

- LDA模型具备不能理解文本词语顺序的问题，例如“我爱你”和”你爱我”，LDA模型认为其意思一致。这也是限制其预测准确度的原因之一。相比较，之后学习的循环神经网络等解决了此类问题，可以在之后的大作业中验证该问题的重要程度。

6.参考文献

1. `https://zhuanlan.zhihu.com/p/106980996`