LDA 模型

ZY2203810

吴金旺

Abstract

本文简单介绍了 Latent Dirichlet Allocation (LDA)模型，并用它对金庸小说进行分类。

Introduction

Latent Dirichlet Allocation (LDA)模型，也被称为 LDA 主题模型。在 LDA 主题模型下，每个词，每段话，乃至篇文章都有自己的主题。它认为每篇文章可以由一定的主题分布来表示，而这一主题分布又是由组成文章、段落的各种词汇直接决定。假设每篇文章的侧重点不同，那么它们将具有不同的主题分布，如果我们从中抽取一些段落，利用段落的词汇求出它的主题分布，将可以利用主题对段落进行聚类。

以上是通俗的说法，具体来说，在用 LDA 模型生成一篇文章i时，我们通过一个狄利克雷分布α抽样得到它的主题分布θ_i，从这一主题分布中继续抽样得到它的第j个词的主题$z_{i,j}$。实际上，一个主题下，不同词语出现的概率不同，因此我们还需用到另一个狄利克雷分布β，抽样得到主题$z_{i,j}$的词语分布$\phi_{z_{i,j}}$，然后从这一词语分布中得到最终的词语$w_{i,j}$。

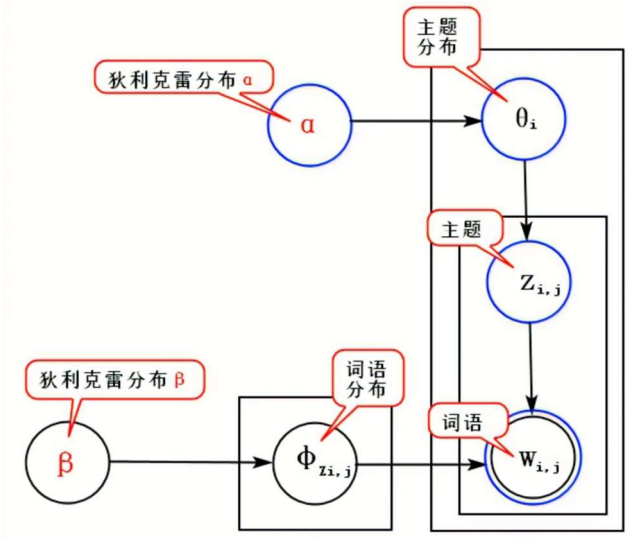


图 1 LDA 模型示意图

那么反过来，我们先统计段落每个词语，得到用词频模拟的词语分布，这一分布就可以用于求出段落的主题，进而实现聚类效果。

Methodology

我们首先对文本进行预处理。由于要求有 200 个具有 500 词的段落，而单一自然段一般较短，我们要先进行分段。简单来说，从每篇文章开始计算，遍历每一个自然段，将它们以此放入一个队列之中，一旦队列长度大于 1500（我们认为一个词语长度应该大约在 2-3 个字，这可以保证分词后长度大于 500），就把队列作为一个段落保存起来。对于每篇文章的最后一个段落，如果长度不足 1500，则直接舍弃。

然后对每一段文本进行过滤，去除所有英文字符和标点符号。按字建模时，直接用空格将每个字分割开，按词建模时，则使用 jieba 分词，且会去除停用词。

我们采用 sklearn 模型库提供的 LDA 模型。首先将每个词向量化，然后送入预设好的 LDA 模型中进行训练，得到每个段落的主题分布，最后用 pyLDAvis 库进行可视化。

Experimental Studies

A.数据集

我们首先抽取数据集，最初的设想是从所有段落组合中随机抽取 200 个，但是仅从文本的字节数目就可以判断出，不同小说的长度相差可能达到几十倍，这样随机抽样，可能字数较少的小说，如《越女剑》就抽取不到或者只抽到一段。为此我们设计了一种平均抽样方法，保证来自不同小说的段落数目相差不会太大（但是代码中也提供了随机抽样方法）。

然后我们就按照上一节中提到的方法，进行分词。每个段落以一个字符串的形式存储起来，字符串中每个词/字之间都有空格，这取决于分词模式。

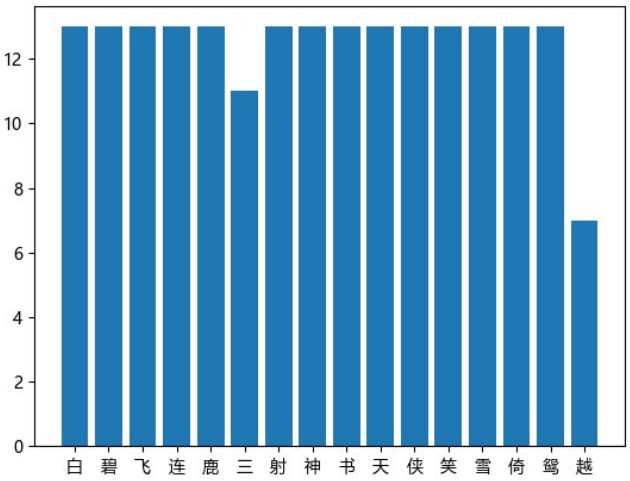


图 2 数据集分布

B.运行结果

我们先考虑用 sklearn 自带的困惑度来求出最佳主题数目。我们从文本库中抽取 100 个段落，并计算它们的适应度，如下图所示。但是显然结果存在很大问题。如果不考虑仅有一个主题的情况，那么困惑度是在不断上升的，这与我们的初衷相违背。

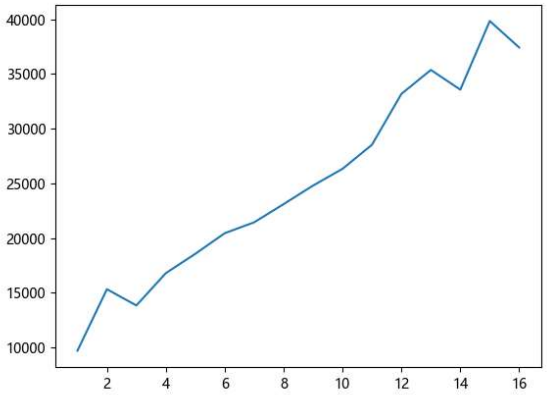
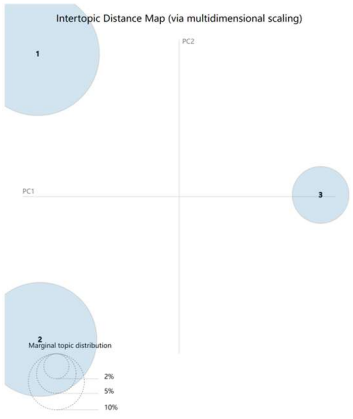


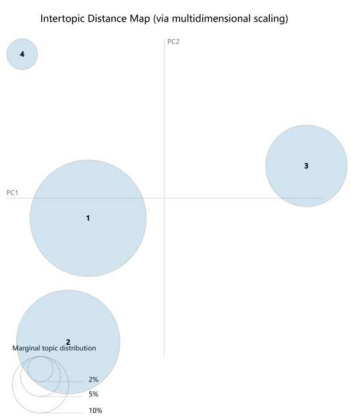
图 3 主题数-困惑度示意图

因此我们还是利用可视化的手段，进行手动挑选。如下图所示，可以看到在我们的数据集下，大概 4-6 个主题是比较合适的。我们选择 4 个主题。

(a) topic=2

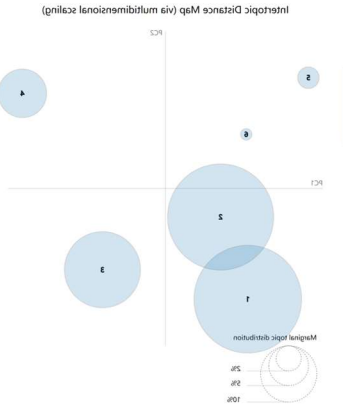


(b) topic=3



(c) topic=4

(d) topic=5

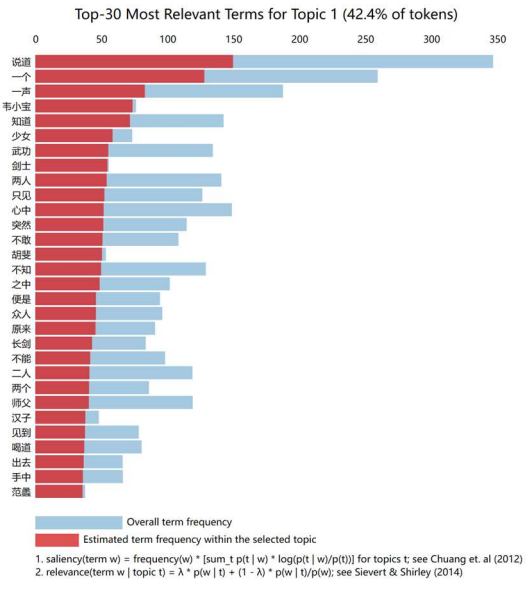


(e) topic=6

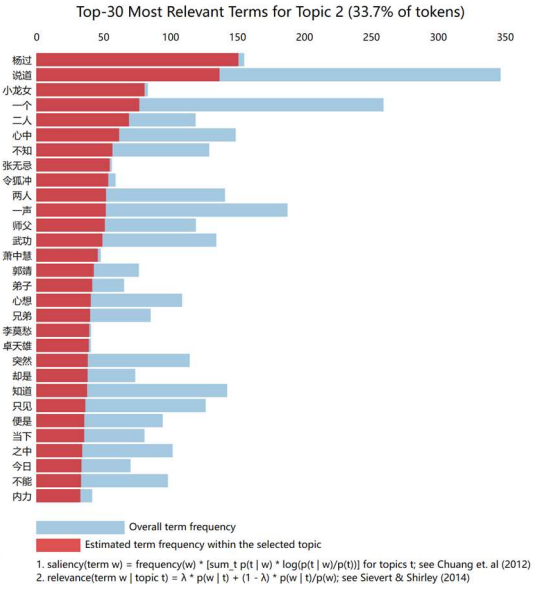
(f) topic=7

图 4 主题-分类示意图

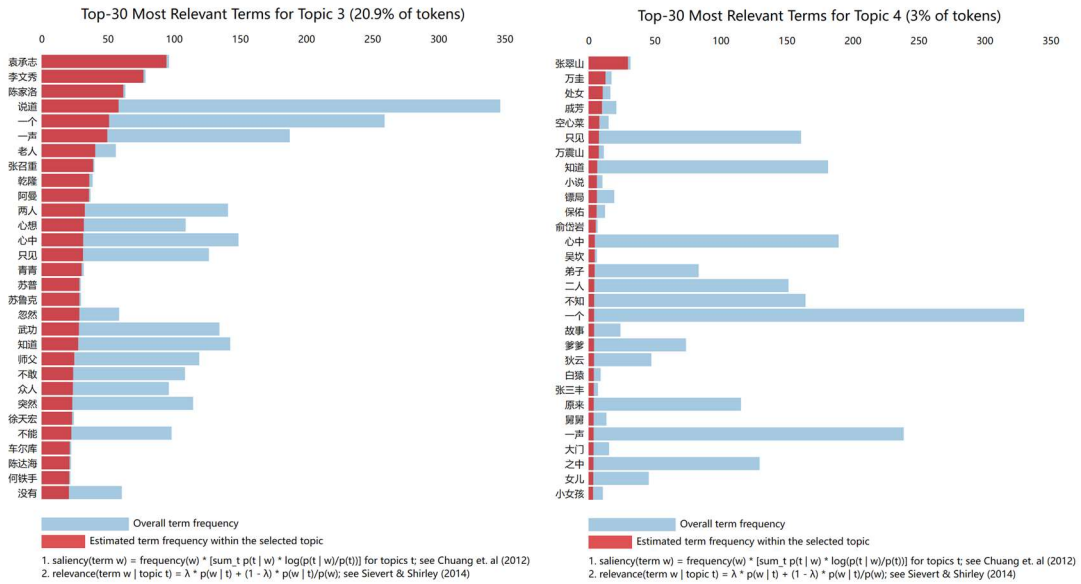
此条件下，四个主题的词频分布主要是：



(a) topic1



(b) topic2



(c) topic3　　(d) topic4

图 5 每个主题的词汇分布图

可以看出，其中包含相当多的人名，比如 topic2 中，包含了杨过和小龙女，张无忌和令狐冲，同一个故事中的人物确实被分到了同一主题。越是专有名词，它越专属于某一个主题，反之，“一个”、“说道”这种词语则分属于各个主题。

下图统计了每本小说的主题统计，以训练集中的段落作为数据求出它的主题分布，以概率最高的主题作为该段落的主题。可以看到，主题和文章之间是存在明显的倾向的。

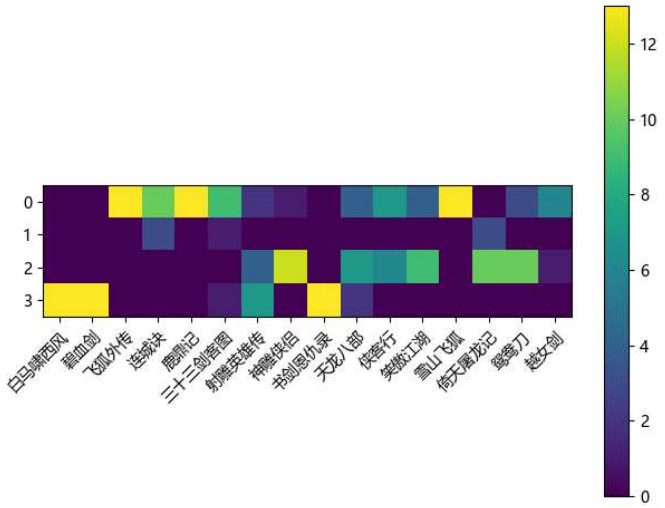


图 6 预测结果分布

最后我们按字进行分类，仍然取 4 个主题，如下图所示。

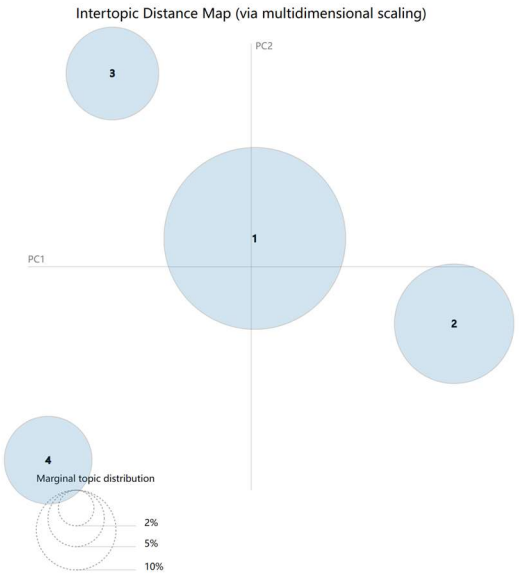
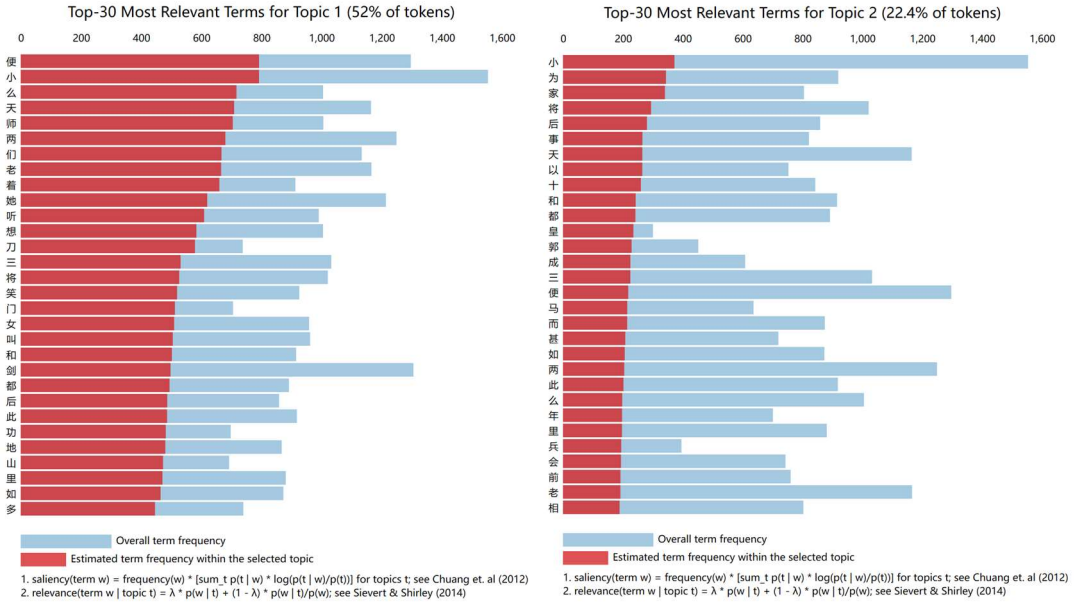
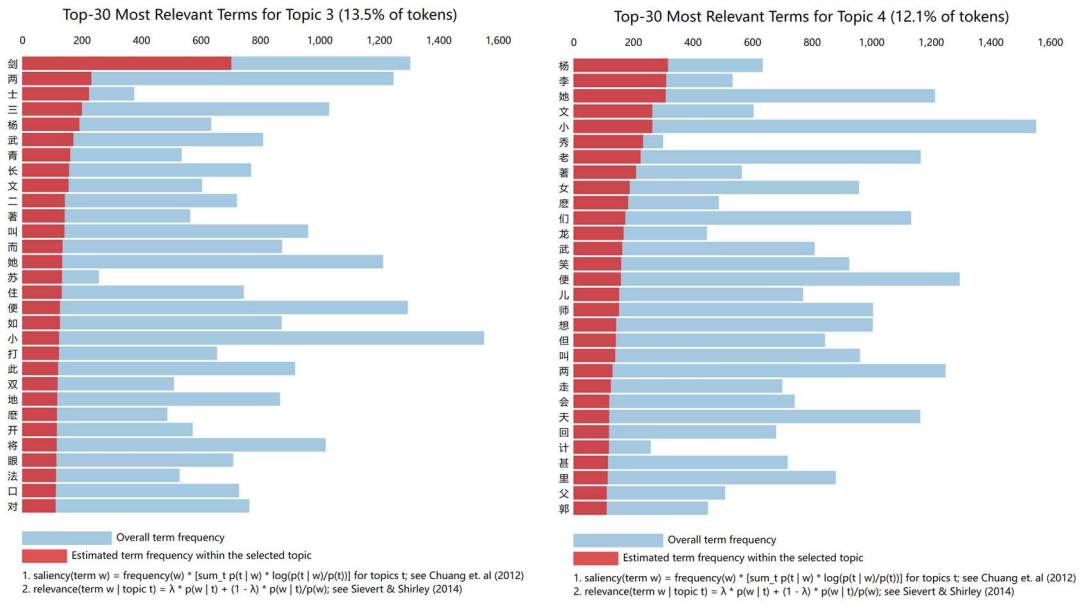


图 7 按字计算的主题-分类示意图

每个主题的词语分布如下：



(a) topic1 (b)topic2



(b) topic3　　(d) topic4

图 7 按字计算的各主题词汇分布图

可以看出，出现频率最多的是组成名字的字和各种常用字。同样的前者广泛分布在各个主题中，后者仅在部分主题分布广泛。同样的，按字计算时，每本小说往往只出现在一个或两个主题中。

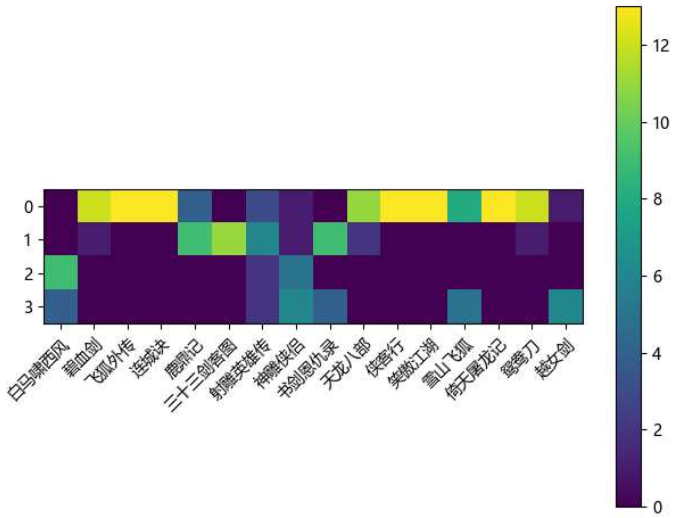


图 6 按字计算的预测结果分布

Conclusion

本次作业分析了使用了 LDA 主题模型对金庸小说进行了主题分类，可以看出，LDA 模型可以无监督构建出文本的主题，是一种有效的模型。