深度学习与自然语言处理第三次作业报告

张甫成
sy2206303@buaa.edu.cn

摘要

本文是深度学习与自然语言处理第三次作业的实验报告，首先介绍了 LDA 模型的原理和建模过程；然后利用 LDA 模型对金庸小说的每个段落进行建模，并把每个段落表示为主题分布后进行分类；最后尝试不同的主题数量，并探究了以"词"和以"字"为基本单元时分类的情况。

绪论

LDA 模型概述

LDA 模型全称 Latent Dirichlet Allocation，是一种主题模型。LDA 模型将一个文档看作是一些主题的集合，每个主题又可以被看做是一些单词的集合。通过 LDA 模型，我们可以得到每个主题对应的单词，以及每个文档中包含哪些主题。LDA 模型在许多领域都有应用，例如，在新闻分类中，可以使用 LDA 模型将新闻归入不同的主题；在社交媒体中，可以使用 LDA 模型来分析用户的兴趣和观点；在市场研究中，可以使用 LDA 模型来分析用户的反馈和评论。

LDA 模型计算方法

LDA 模型的目标是通过给定文档中的单词，推断每个文档中的主题分布，以及每个主题中的单词分布。为了实现这一目标，LDA 模型使用了 Gibbs 采样算法或 EM 算法进行迭代计算。Gibbs 采样算法是一种基于马尔可夫链蒙特卡罗(MCMC)方法的抽样算法，它可以从联合分布中抽样得到样本，从而计算出联合分布的期望值和方差等统计量。在 LDA 模型中，Gibbs 采样算法可以从主题分布和单词分布中抽样得到主题和单词。

具体来说，假设我们有一段文本d，其中包含N_d个 词$w_{d,n}$，并且我们已经推断出了文档d中每个单词的主题$z_{d,n}$和每个主题中的单词分布$\phi_{k,v}$和每个文档中的主题分布θ_d，我们可以用一下的过程从这些分布中抽样得到新的主题$z_{d,n}^{new}$和单词$w_{d,n}^{new}$：

1. 对于文档d中的每个单词$w_{d,n}$，先将其主题$z_{d,n}$的计数减一，同时将每个主题中单词$w_{d,n}$的计数减一。
2. 根据每个单词的主题分布$p(z_{d,n}=k|w_{d,n}=v,\mathbf{z}-d,n,\mathbf{w}d,-n,\boldsymbol{\theta}_d,\boldsymbol{\phi}k)$，重新计算每个单词的主题$zd,n$。
3. 根据每个主题中的单词分布$p(w_{d,n}=v|z_{d,n}=k,\mathbf{z}-d,n,\mathbf{w}d,-n,\boldsymbol{\theta}_d,\boldsymbol{\phi}k)$，重新计算每个单词的单词$w_{d,n}$。

重复执行步骤 1 到步骤 3，直到达到一定的抽样次数，即可完成建模。

随机森林

随机森林（Random Forest）是一种集成学习算法，它通过构建多棵决策树并结合它们

的预测结果来提高预测性能。随机森林中的每棵树都是独立构建的，它们之间没有依赖关系。

随机森林算法在构建每棵树时都会引入随机性。具体来说，它会对训练数据进行有放回抽样，以生成不同的训练子集；此外，在选择分裂特征时，它会从所有特征中随机选择一部分特征，然后从中选择最佳分裂特征。这些随机化策略有助于降低模型的方差，提高模型的泛化能力。

在预测时，随机森林会将所有树的预测结果进行汇总。对于分类问题，它会采用投票的方式，选择票数最多的类别作为最终预测结果；对于回归问题，它会计算所有树的预测值的平均值作为最终预测结果。

随机森林算法具有很好的可扩展性和鲁棒性，可以处理高维数据和大量噪声数据。由于 LDA 模型给出的主题分布很难线性可分，因此 k-means 等分类器失效，即使用 kernel-SVM 也达不到较高的准确率，因此使用随机森林作为分类器，对主题分布进行分类。

方法

我们首先介绍分割语料库的方法，然后介绍模型的构造方法，最后模型的评价方法。

token 序列的获取方法

我们首先将语料库按照换行符进行分割，得到段落，之后将每个段落分割为字序列或词序列，删除标点和停用词后，去除长度小于 500 的段落。

如果总段落数大于 200，则通过设定切片步长，将段落数量降低到 200 附近。

LDA 模型的构造方法

我们使用 gensim 库建立 LDA 模型。首先遍历主题数量，计算所得模型的困惑度和一致性，然后选取困惑度最低、一致性最高的主题数量进行实际建模。

分类模型

得到 LDA 模型后，即可将一段文本转换为一个 N_topics 维的向量。我们按照此方法将 token 序列转换为主题分布后，划分出 90%的序列作为训练集，训练随机森林分类器，最后计算该分类器在剩余 10%的测试集上的 top1 准确率、top5 准确率。

实验结果

分割后的数据量

按照词分的数据量如图 1 所示。

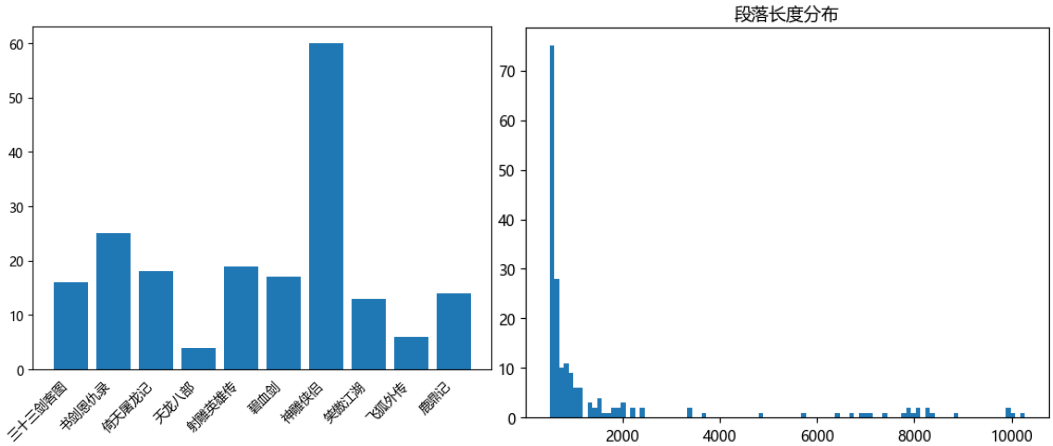

图 1: 按照词分，每部小说段落数和段落长度分布

按照字分的数据量如图 2 所示。

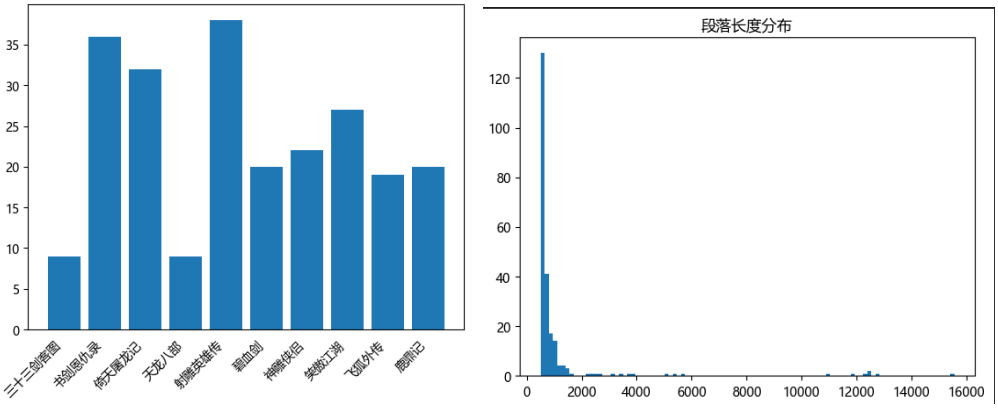

图 2: 按照字分，每部小说段落数和段落长度分布

N_topics 的选取

模型困惑度的对数、一致性随着 N_topics 的变化如图 3、图 4 所示，其中，图 3 是以"词"为基本单元的结果，图 4 是以"字"为基本单元的结果。根据结果，将词模型的 N_topics 定为 10，将字模型的 N_topics 定为 80。

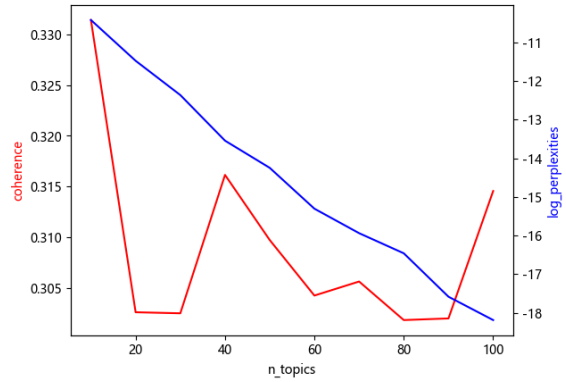

图 3：按词分的 LDA 模型质量

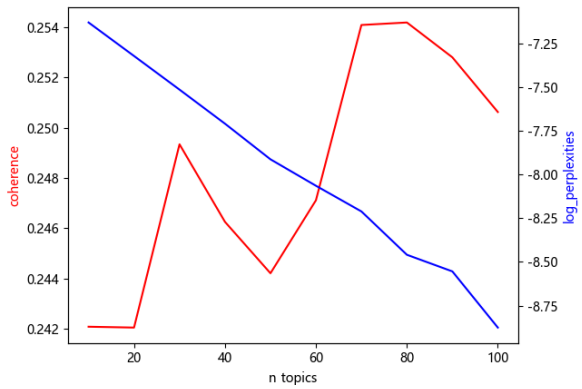

图 4：按字分的 LDA 模型质量

分类结果

我们的方法分类结果如表 1 所示。

	Top1 准确率	Top5 准确率
字序列	33.3%	87.5%
词序列	40.0%	90.0%

结论

我们分别用字序列和词序列对金庸小说的段落进行 LDA 建模，得出了分词比分字更适合 LDA 建模的结论。在实验过程中，发现模型的困惑度随着 N_topics 的上升而稳定下降，但一致性的变化不稳定，可能是由于语料库规模较小。分类器给出的分类结果随机性较大，未来可以对金庸小说语料库进行更均匀的划分，增加参与 LDA 训练的数据规模，平衡序列的长度和不同小说的序列量，以实现更稳定的建模。