基于 LDA 模型的主题建模

摘要

本文通过对 16 本中文小说进行均匀抽取 200 个段落（每个段落大于 500 个字），每个段落的标签为其所属小说，利用 LDA 模型对文本进行建模，将每个段落表示为主题分布后进行分类。实验分析了不同主题数下分类性能的变化，并比较了以"词"和以"字"为基本单元下分类结果的差异。实验结果表明，选择适当的主题数对分类性能有显著影响，以"字"为基本单元的分类结果略优于以"词"为基本单元的结果。

关键字：LDA 模型；主题建模

一、 引言

LDA（Latent Dirichlet Allocation，隐含狄利克雷分布）是一种用于主题建模的概率生成模型，由 Blei、Ng 和 Jordan 在 2003 年提出。在信息检索、自然语言处理、社交媒体分析等领域有广泛应用。随着中文自然语言处理技术的发展，中文文本分类已经成为了一个热门的研究领域。然而，中文文本分类与英文文本分类有很大的不同，其中最重要的就是中文的基本单元是"字"而非"词"，因此需要对中文文本分类进行一定的调整。

二、 LDA 建模

LDA 模型是 2003 年由 David Blei, Andrew Ng 和 Michael I. Jordan 提出的一种主题模型，是一个三层贝叶斯概率模型，包含词、主题和文档三层结构。其基本思想是，每篇文档中都含有若干个主题，每个主题又由若干个词语组成，文档中每一个词都由其中的一个主题生成。在 LDA 模型中，主题是隐含变量，词语是观测变量。LDA 假设文档的主题是通过从一个 Dirichlet 中采样得到的，而每个主题又是从另一个 Dirichlet 中采样得到的。具体来说，LDA 模型包含以下几个基本元素：

1) 文档集合：包含多篇文档，每篇文档由多个词语组成。
2) 词典：所有文档中出现的不同词语的集合。
3) 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。
4) 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 模型的训练过程是一个迭代的过程，具体步骤如下：

1) 初始化每个词语的主题，可以随机初始化或根据先验知识初始化。
2) 遍历每篇文档中的每个词语，计算每个词语属于每个主题的概率分布。
3) 根据计算得到的概率分布，重新分配每个词语的主题。
4) 重复步骤 2 和 3，直到主题分布收敛为止。

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、推荐系统等应用。

三、 实验与分析

1. 数据集

本实验使用了 16 本中文小说作为数据集，从中随机抽取 200 个段落，每个段落的长度均大于 500 字。

2. 数据预处理

对于中文文本，需要进行分词、过滤停用词等预处理工作。本实验使用了 jieba 库对中文文本进行分词，使用中文停用词表对文本进行停用词过滤。同时，由于本实验旨在比较以"词"和以"字"为基本单元下分类结果的差异，因此对于以"词"为基本单元的情况，保留了中文文本的分词结果，而对于以"字"为基本单元的情况，则将中文文本拆分为单个字。

此外，构造一个词典，为文档中的每个词分配一个独一无二的整数编号，然后统计各词的词频。

3. LDA 模型训练

LDA 是一种基于贝叶斯概率模型的主题建模方法。它将每个文档数据看作是由若干个主题和主题下的词语构成的 Dirichlet 分布，其中每个主题对应一些词语的概率分布 $\text{Dirichlet}(\beta), \beta = (\beta_1, \beta_2, \ldots, \beta_N)$（$N$表示每个主题中词语的数量），每个文本则包含若干个主题的混合，即$\text{Dirichlet}(\alpha), \alpha = (\alpha_1, \alpha_2, \ldots, \alpha_k)$（$k$表示主题数量），从而实现对文本的主题建模。LDA 模型包含三个随机过程：文档生成主题，主题生成单词，以及文档生成单词。通过使用 Gibbs 抽样算法进行参数估计α, β，可以得到每个文本的主题分布和每个主题下词语的分布概率。

本文使用 gensim 库中的 LdaModel 函数进行模型训练。在模型训练过程中，需要选择合适的主题数。本文分别比较了基于字和词的 LDA 模型的主题数从 2 到 29 的情况下困惑度、主题一致性的变化，如图 1 所示。

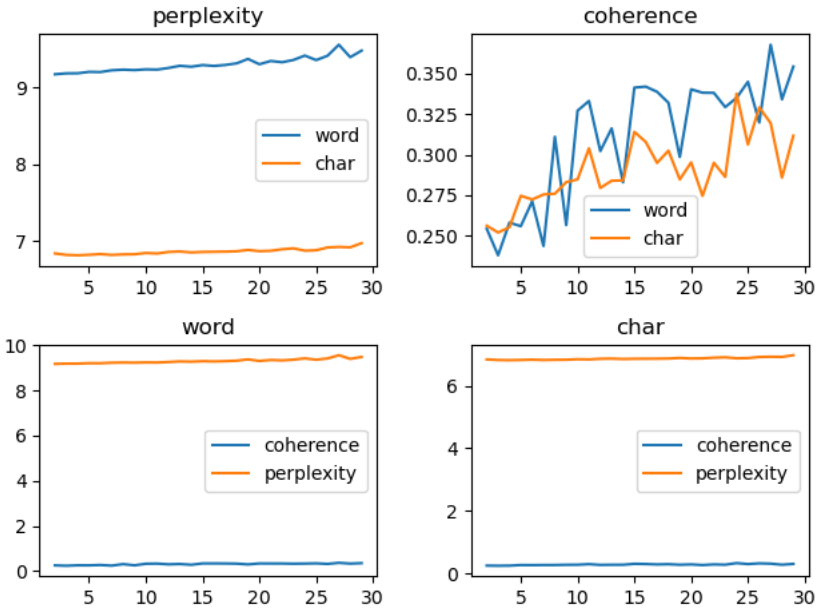

图 1 LDA 模型的困惑度和主题一致性

4. 实验结果分析

困惑度表示模型对未见过的文本数据的预测性能。困惑度越小，模型的预测性能越好。主题一致性是通过计算每个主题中单词之间的相似度来评估主题的质量。主题一致性的值越大，主题之间的相似度越高，LDA 模型的质量越好。

在基本单元不同的情况下，LDA 模型对文本建模的效果可能会存在差异。从图 1 观察发现，基于字（char）的 LDA 模型的困惑度比基于词(word)的低，主题一致性的波动不显著。因此，以“字”为基本单元进行 LDA 模型的训练，可能会更准确地表达文本的含义。

从图 1 观察发现，随着主题数的增加，主题一致性增大，每个主题中单词之间的相似度越高，主题的连贯性越好。

综上所述，以“字”作为基本单元的 LDA 模型在对中文文本进行建模可能会更加准确和细致，但同时也需要注意过拟合问题。以“词”作为基本单元的 LDA 模型则可能会更加关注语义信息，但可能会忽略一些细节信息。因此，在具体应用中需要根据任务需求和文本特点来选择合适的基本单元。

四、 总结

本文通过 LDA 模型对中文小说进行主题建模，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。