Report of Deep Learning for Natural Language Processing

种雨萌

2775965921@qq.com

Abstract

本篇报告从金庸的 16 部小说中均匀抽取 1000 个段落作为数据集，通过 LDA 模型进行文本建模，并把每个段落表示为主体分布后进行分类。讨论了不同主题个数，以“字”和“词”为基本单元，以及长短文本对分类性能和结果的影响。

Introduction

LDA，即 Latent Dirichlet Allocation（潜在狄利克雷分配），是一种常用的文本主题模型，它通过分析一些文档抽取出它们的主题分布，对其进行主题聚类或文本分类。LDA 被广泛应用于文本挖掘、信息检索和自然语言处理等领域。本次实验基于 LDA 模型完成以下任务。

从链接给定的语料库中均匀抽取 1000 个段落作为数据集（每个段落可以有 K 个 token，K 可以取 20，100，500，1000，3000），每个段落的标签就是对应段落所属的小说。利用 LDA 模型在给定的语料库上进行文本建模，主题数量为 T，并把每个段落表示为主题分布后进行分类（分类器自由选择），分类结果使用 10 次交叉验证（i.e. 900 做训练，剩余 100 做测试循环十次）。实现和讨论如下的方面：

（1）在设定不同的主题个数 T 的情况下，分类性能是否有变化？

（2）以“词”和以“字”为基本单元下分类结果有什么差异？

（3）不同的取值的 K 的短文本和长文本，主题模型性能上是否有差异？

Methodology

M1: 数据集构建

在抽取段落前，需要对文本进行预处理。首先去除文本中的停用词，若以词为单位，还需使用 jieba 库进行分词。

由于每篇小说字数不同，从中均匀抽取 1000 个段落需要统计每篇小说的长度，按照比例分配每篇小说应抽取的段落数量，确保选择不重复的小说段落，并将段落编号、文章标签和段落内容等信息保存下来。

M2: LDA 模型

LDA 的核心思想是假设文档集合中的每个文档都是由多个主题混合而成的，而每个主题又对应一个词的分布。换句话说，LDA 认为文档是由主题的混合构成的，而每个主题又对应一组词，文档中的每个词都是由这些主题按一定概率生成的。LDA 模型做出了以下两个关键假设：

（1）文档-主题分布假设。每个文档在主题上的分布是由一个参数为θ的狄利克雷分布生成的，表示文档中各个主题的比重。

（2）主题-词分布假设。每个主题在词上的分布是由一个参数为β的狄利克雷分布生成的，表示每个主题中词的分布。

基于这两个假设，LDA 模型通过对文档集合中的词进行观察和统计，推断出文档的主题分布和每个主题中词的分布。具体而言，LDA 模型使用了贝叶斯推断的方法，通过吉布斯采样等算法来估计参数θ和β，从而实现对文本数据的主题建模。

LDA 模型的生成过程包括以下三个步骤：

（1）对一篇文档的每个位置，从主题分布中抽取一个主题；每个文档的主题分布遵循一个狄利克雷分布。

（2）从上述被抽到的主题所对应的单词分布中抽取一个单词；这个分布也遵循一个狄利克雷分布。

（3）重复上述过程直至遍历文档中的每一个单词。

M3：支持向量机分类器

支持向量机（Support Vector Machine，SVM）是一种常用的监督学习算法，主要用于分类和回归任务。支持向量机分类器（Support Vector Classifier，SVC）是 SVM 的一种变体，被广泛应用于文本分类、图像识别、生物信息学等领域，具有较好的分类性能和鲁棒性。同时，SVC 也具有一定的局限性，例如对大规模数据集的处理效率较低，需要较长的训练时间。

与传统的 SVM 相比，SVC 在处理非线性可分问题时引入了核函数，通过将

输入特征映射到高维空间中，从而使得样本在新的空间中线性可分。常用的核函数包括线性核、多项式核、径向基函数核等。本次实验中，设定 SVC 核函数为线性核。

Experimental Studies

本实验设置 LDA 模型的主题个数 T 依次为 10、20、50 和 100，段落长度 K 依次为 20、100、500、1000 和 3000。在以字为单位时得到分类平均准确率如下表所示：

T \ K	20	100	500	1000	3000
10	0.166	0.253	0.571	0.673	0.792
20	0.126	0.287	0.678	0.730	0.872
50	0.129	0.226	0.659	0.781	0.924
100	0.116	0.219	0.637	0.793	0.914

在以词为单位时得到分类平均准确率如下表所示：

T \ K	20	100	500	1000	3000
10	0.083	0.184	0.304	0.378	0.604
20	0.127	0.169	0.314	0.434	0.730
50	0.114	0.153	0.291	0.546	0.738
100	0.116	0.131	0.347	0.516	0.790

根据实验结果可以总结出以下规律。

（1）设定不同的主题个数 T，会影响到分类性能。在长文本（K 值较大）的情况下，T 越大，分类准确率越高，分类器的表达能力和拟合能力越好；在短文本（K 值较小）的情况下，T 越大，准确率反而下降，这可能是因为短文本包含的信息和语义内容较少，主题数量增加可能会导致特征空间更加稀疏，造成模型过拟合，使得准确率下降。

（2）以“词”和以“字”为基本单元，分类结果差异较大。在同样的 T、K 取值下，以字为单位的分类准确率总是高于以词为单位。这可能是因为以字为单位的主题分布更能反映各小说的语言风格和文学特征，使得分类性能提高。

（3）不同的取值 K 的短文本和长文本，主题模型性能上存在一定差异。长文本通常包含更多的语义信息，不同主题之间更容易区分和识别，更适合使用 LDA 模型处理；短文本篇幅有限，且更容易受到噪声的影响，可能导致主题模型的性能较差。这就导致短文本的分类准确率总是低于长文本。

Conclusions

本实验基于 LDA 模型，实现了金庸小说的语段分类任务。首先对小说文本进行预处理，去除停用词并使用 jieba 库进行分词。接着统计每部小说的长度并按照比例分配每篇小说需抽取的段落数量。然后通过 LDA 模型训练得到主题特征向量。最后采用 SVC 分类器对段落进行分类，并通过 10 次交叉验证评估分类器的平均准确率。基于以上步骤，本实验研究和讨论了不同主题个数 T、不同基本单元（字或词）以及不同段落长度 K 对分类性能的影响。实验结果表明，T、K 越大，分类效果越好；以字为单位的分类性能普遍高于以词为单位；LDA 模型对长文本的处理能力较强。