Deep Learning and Natural Langauge Processing

——第二次作业

卢田雨 ZY2303525
2589402656@qq.com

Abstract

本文从给定的语料库中均匀抽取 1000 个段落作为数据集，选定每个段落的标签为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，把每个段落表示为主题分布后进行分类，分类结果使用 10 次交叉验证。研究了不同主题数下对分类性能的影响，以"词"和以"字"为基本单元下分类结果有什么差异，不同的取值的 token 数量的短文本和长文本对主题模型性能上是否有差异。

Introduction

LDA（Latent Dirichlet Allocation）模型是一种概率主题模型，常常用来文本分类。它最早由 Blei David M.，Ng, Andrew Y.，Jordan 等人在 2003 年提出，旨在推测文档的主题分布。它可以将文档集中每篇文档的主题以概率分布的形式给出，从而通过分析一些文档抽取出它们的主题分布后，便可以根据主题分布进行主题聚类或文本分类。基于 LDA 模型，本次实验将要研究以下几个问题。

从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）。实现和讨论如下问题：

（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？

Experimental Studies

LDA 模型参数主题数量 T 取值为 10、20、50、100，tkoen 数 K 取值为 20、100、500、1000。并分别对以词为单位和以字为单位进行分类试验。所得结果如下：

表 1：基于以字为单位的验证集的分类器平均准确度

T \ K	20	100	500	1000
10	0.116	0.238	0.541	0.673
20	0.124	0.295	0.708	0.803
50	0.125	0.319	0.760	0.880
100	0.132	0.357	0.765	0.895

表 2：基于以词为单位的验证集的分类器平均准确度

T \ K	20	100	500	1000
10	0.114	0.179	0.308	0.394
20	0.121	0.162	0.335	0.512
50	0.124	0.145	0.321	0.651
100	0.144	0.155	0.415	0.744

（1）在设定不同的主题个数 T 而其他情况不变的情况下，主题数 T 越大，分类器的性能越好，准确度越高。这表明，当 T 增大时，分类标准越细化，各种语义词与主题有很好的对应，分类器正确分类的可能性越大。

（2）以“词”和“字”为基本单元的分类结果差异较大。由实验数据看出，以字为单元的分类性能高于以词为单元。但随着 K 的增大，两者的差异性逐渐缩小。

（3）不同的取值的 K 的短文本和长文本， K 越大分类器的分类性能越好。这表明，段落越长，分类器能够从上下文获取到的更丰富的语义词特征，可以有效提升分类器分类性能。