Report of NLP Assignment 2

Weidong Xu
2697782204@qq.com

Abstract

实验要求：以文学类作品（金庸小说）为对象，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，并依次涵盖设定不同的主题个数T、以“词”和“字”为基本单元下分类和不同分类器性能比较来展开实验。

Introduction

主题分析模型（Topic Model）是以非监督学习的方式对文档的隐含语义结构进行统计聚类，用以挖掘文本蕴含的语义结构的技术。隐含狄利克雷分布（Latent Dirichlet Allocation, 简称 LDA）是常用的主题模型计算方法，是基于贝叶斯思想的无监督的聚类算法，广泛用于文本聚类，文本分析，文本关键词等场景。LDA主题模型主要用于推测文档的主题分布，可以将文档集中每篇文档的主题以概率分布的形式给出根据主题进行主题聚类或文本分类。

LDA主题模型不关心文档中单词的顺序，通常使用词袋特征（bag-of-word feature）来代表文档。

Experimental Studies

本实验聚焦于以下四个方面，分别进行探讨:

1. 不同主题数量T对分类性能的影响：不同主题数量对分类性能影响不一。

2. 以“字”或“词”为单位，分别评估对LDA模型的影响是否显著。

3. 文本长度K值对模型性能预期的影响。

4. 分类器对模型的性能差异。本实验使用catBoost作为分类器，其优点在于对分类型特征的处理。这使得在训练模型之前可以考虑不用再通过特征工程去处理分类型特征，同时借助困惑度（perplexity）来进一步探讨模型性能的好坏。

具体测试数值如下表所示：

clif_model	K value	splitByword	n_topics(个)	Accuracy(mean+-std)(%)
catBoost	3000	FALSE	100	0.87+-0.05
catBoost	3000	FALSE	50	0.70 (+/- 0.16)
catBoost	3000	FALSE	20	0.48 (+/- 0.11)
catBoost	1000	FALSE	100	0.64 (+/- 0.17)
catBoost	500	FALSE	100	0.45 (+/- 0.14)
catBoost	100	FALSE	100	0.17 (+/- 0.07)
catBoost	20	FALSE	100	0.14 (+/- 0.03)
catBoost	3000	TRUE	100	0.27 (+/- 0.12)

para_len对主题模型性能影响	para_len（个）		
splitByword		n_topics(个)	perplexity
FALSE	500	20	1.1669E+13
FALSE	100	20	4.08873E+14
FALSE	20	20	3.11167E+15

此外，本实验为了测试分类器对模型的性能差异，还引入了随机森林和K近邻算法进行比对，得出结论：主题数越多，分类性能越好；以词划分，分类性能更好；文本长度K取值越大，分类效果越好，LDA模型训练时间越长；引入困惑度指标，K取值越大，困惑度越低，LDA模型性能越好。