Central University of Finance and Economics

中央财经大学

大数据分析统计基础课程

第七次课作业

吴宇翀

2021210793

email@wuyuchong.com

指导老师：盖玉洁

2021 年 11 月 8 日

目录

1 摘要

我们使用 IMDB 数据集进行文本分类。在文本预处理阶段，我们尝试使用词编码和词向量的方式，在训练阶段，我们构建了 DNN、LSTM、BERT 等多个深度学习模型进行训练，并进行了模型比较，最高达到了 99% 的准确率。最后，为了进一步实现在超大文本集上进行训练，我们使用基于 Spark 的分布式算法在集群服务器上进行训练测试。

模型	计算配置	用时	准确率	可拓展性
tokenize + DNN	阿里云服务器 Xeon 8 核 CPU 32G 内存	10 分钟	60%	低-单机
Word2Vec + LSTM	阿里云服务器 Xeon 8 核 CPU 32G 内存	2 小时	80%	低-单机
bert - 小型	阿里云服务器 Xeon 8 核 CPU 32G 内存	1 小时	90%	低-单机
bert - AL	阿里云服务器 Xeon 8 核 CPU 32G 内存	1.5 小时	90%	低-单机
bert - 标准	阿里云服务器 Xeon 8 核 CPU 32G 内存	3 小时	92%	低-单机
spark	中央财经大学大数据高性能分布式集群	– 分钟	-%	高-集群

分布式模型在该小型数据集上没有优势，进行此项的意义在于对大型文本数据集可拓展性的技术储备，仅有在文本量级超过单机可承载上限时，分布式计算才具备意义