2018 网络零售平台商品分类项目

概要介绍

团队：TempName 小组

目录

1. 前言

1.1 编写目的

本说明书目的在于明确说明项目各功能的实现方式，指导开发员进行编码。

本说明书的预期读者为：系统设计者、系统开发员。

1.2 背景

在我们的生活当中可能会有类似的经历，我们想要购买某件商品，一般是搜索该商品的类别去查询我们想要的产品，但是会存在着搜出来的商品和我们想要商品的类别不符的情况，这就是由于分类不精确导致的问题。因此我们的项目旨在针对来自不同零售平台的商品，能够通过其商品描述信息自动高效地判定其类别。

1.3 术语

- **商品标签：**在这里特指商品的大类（体现商品生产和流通领域的行业分工，如珠宝饰品）、中类（体现具有若干共同性质或特征商品的总称，如翡翠玉石）和小类（对中类商品的进一步划分，体现具体的商品名称，如项链）的组合。
- **分类：**指利用 50 万个商品包含的标签信息，对剩余的 450 万个商品进行合理的标签判定。

2. 创意描述

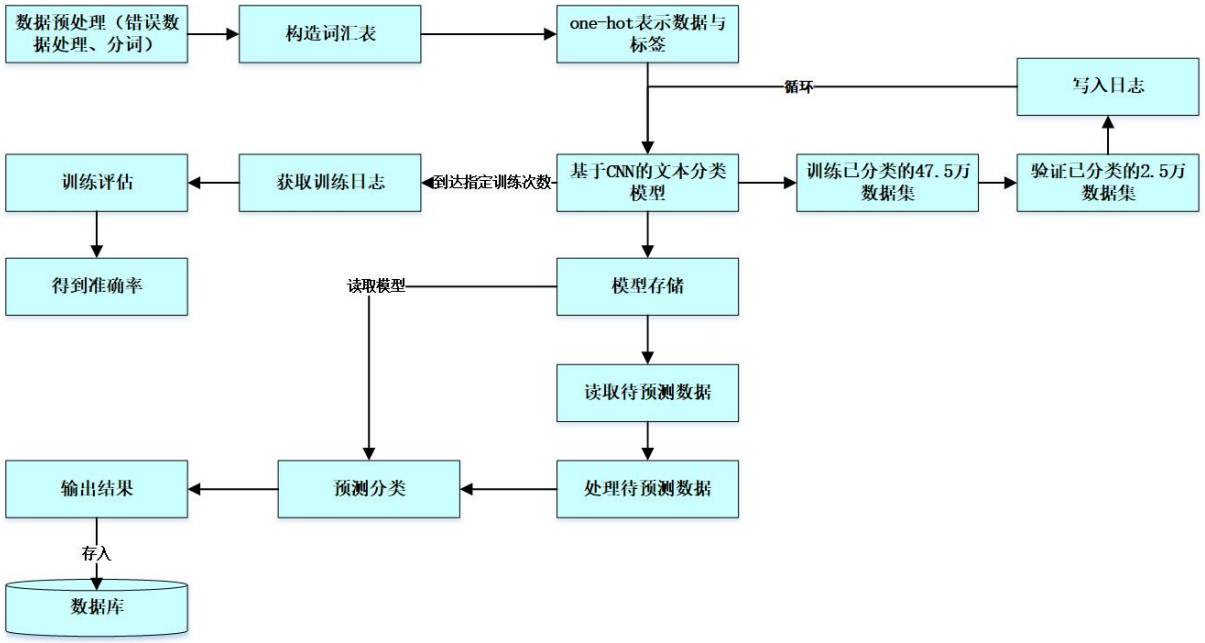
本项目采用基于 CNN 的文本分类模型实现自动分类。文本分类模型大体上分为基于传统机器学习和基于深度学习的文本分类模型，后者与前者最主要的区别是随着数据规模的增加其性能也不断增长。本项目的数据集在万级以上，因此基于深度学习的文本分类模型能够更加完美地解释它。随着现在大数据时代的到来，基于深度学习模型的文本分类模型已经成为了主流，其中 CNN 模型在文本分类任务中是兼具效率与质量的理想模型。因此基于 CNN 的文本分类模型具有良好的商业价值和社会应用价值。

3. 功能简介

本项目的业务需求是通过对 50 万带有商品分类标签的商品进行训练，建立分类模型，对 450 万不带分类标签的商品进行分类，将分类结果存入数据库并用 web 端可视化显示出来。

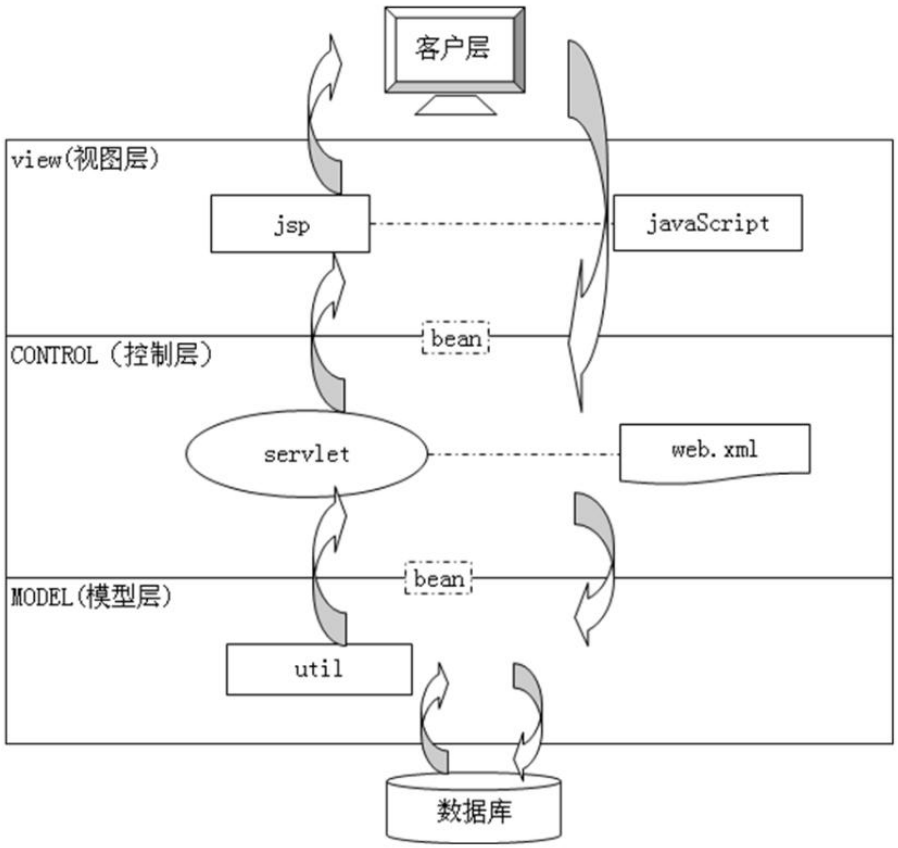
首先对已知的五十万数据进行预处理，（除去非文本内容，处理编码、某些行存在几百条数据堆积等问题）对合法的中文和英文数据进行结巴分词处理并构造词汇表，利用 one-hot 技术来表示数据与标签，训练集和验证集的比例为 19:1，其中 47.5 万数据作为训练集，2.5 万数据作为验证集。采用模型 TextCNN 对我们的数据集进行反复训练和验证，通过反复调试参数选取一个最佳的模型，用已分类的 2.5 万数据与其已知的分类结果进行比对得到准确率，接着读取待分类的 450 万数据同样进行预处理，预测其分

类结果写入数据库。具体流程如图一所示：



图一. 基于 CNN 的文本分类模型具体实现流程图

Web 端采用 MVC 架构，实现整体数据显示、单一查询和批量查询这三个功能。其中整体数据显示以分页表格的形式呈现数据库中所有的数据；单一查询通过搜索商品信息，显示出对应的分类信息；批量查询通过上传文件的方式显示出该文件所有的商品信息以及对应的分类信息。其实现框架如图二所示：



图二. MVC 实现框架

4. 特色综述

1. 采用针对大量数据集的深度学习框架从而可以自动地从已构建的数据集上归纳出一套分类规则；
2. 采用结巴中文分词技术能够将句子最精确地切开，适合文本分析；
3. 采用 One-Hot 技术使文本数值化能够有效降低异常值对模型的影响，增强模型稳定性；
4. 采用目前业界普遍认为准确度最高的模型 TextCNN 进行文本分类，兼具效率与质量；
5. 采用 MVC 架构实现用户与系统之间的交互，支持多种查询数据的方式，可视化效果好。

5. 开发工具与技术

5.1 神经网络中文文本分类：

分类	名称	版本
开发工具	pycharm	5
重要库	jieba（中文分词）	0.39
	pandas	0.24.1
	numpy	1.16.2
	tensorflow	1.13.1
数据库平台	Mysql	5.7.21
代码托管平台	GitHub	1.3.3

5.2 可视化 web 端开发：

分类	名称	版本
开发工具	MyEclipse	4.6.1

应用平台	Tomcat	9.0
开发平台	JDK	1.8.0
数据库平台	Mysql	5.7.21

6. 应用对象

在本项目中主要针对于网上零售平台的商品，实际上只要数据集有文本属性都可以对其分类。

7. 应用环境

7.1 硬件环境

服务器	配置
应用和数据库服务器	内存 64G
	至强 CPU

7.2 软件环境

分类	名称	版本
操作系统	Windows10/Linux	专业版
数据库平台	Mysql	5.7.21
应用平台	Tomcat	9.0
开发平台	JDK	1.8.0

8. 结语

整个模型读取 450 万待预测数据并处理的效率为 12，8571 个/分钟，打标签的效率为 97，825 个/分钟、准确率为 85.792%，为训练集打标签的准确率为 91.593%。实际上如果原始数据集更加规范准确、每类商品信息的数据集更大，我们模型的准确率会更高。

可视化 Web 端的可访问链接:http://www.bestdoublelin.com:8080/fuwu/showdata