YNU 热点问题情感极性分析

——基于云南大学校园集市平台

AngLee

2023 Spring MachineLearning Course
YunNan University

开题报告 4.24

雲南大學
YUNNAN UNIVERSITY

大纲

大纲

雲南大學
YUNNAN UNIVERSITY

校园集市概述

- 云大师生可以通过发帖的形式在校园集市平台进行求助，分享；对于校园热点问题（例如食堂满意度、外卖问题等），大家往往持有不同的意见
- 通过对集市数据的分析，使用恰当的分析框架，就可以判断某段时间学生的心情指数和对校园热点问题的主观倾向

大纲

NLP 情感极性分析

- NLP 即自然语言处理，Sentiment Analysis（情感极性分析）是自然语言处理的一个重要分支，通过分析文本信息，判断个人对事件的主观倾向：积极/消极，实现舆情监控和风险预警
- Sentiment Analysis 经过多年的发展，发展出，数据字典、机器学习、深度学习等多种分析方式

大纲

数据来源

- 数据获取：通过网络爬虫的方式，爬取云南大学校园集市的发帖信息，通过人工标注的方式，构建训练集
- 数据规模：受限于服务器限制，集市平台仅保留近一周的发帖数据，通过数据统计，近 5 天平均日发帖数量为 150 条，以 60 天作为数据获取周期，本项目的数据规模为 9000 条

大纲

模型选择

- 实现 Sentiment Analysis 的方法众多，本次分别使用情感辞典，集成学习和深度学习的方式完成训练
- 采用知网开源的情感辞典 Hownet 进行初步拟合
- 采用集成学习的方式，使用 adaboost 将 SVM 和朴素贝叶斯两种弱分类器集成为强分类器，完成拟合
- 采用 LSTM 深度学习模型进行验证。由于不同词语对情感类别的影响不同，通过添加 attention 机制进行拟合

大纲

雲南大學
YUNNAN UNIVERSITY

模型验证

- 基于 OpenAI gpt turbo 3.5 prompt 的模型验证
- 基于基于百度飞桨 PaddleNLP 的模型验证

大纲

雲南大學
YUNNAN UNIVERSITY

集成学习的使用

- 传统情感分析中，研究方向有传统机器学习领域和深度学习领域的两个方向，对于深度学习方向，新的成果层出不穷，而传统机器学习方向则更多的关注于如何优化单一的分类器以求得更好的效果，鲜有人使用继承学习的方式进行验证。

大纲

Attention 机制的使用

- Attention 机制是 Transformer 的核心思量，而后者则是 GPT 模型的基础，受限于硬件设备和数据集，我们无法复现 GPT 的训练过程，但通过在简单的情感分析任务中引入类似的 Attention 机制，完成分类任务，是对其内核的一次简单实践。

大纲

调整 prompt 验证模型

- 通过调用 OpenAI gpt turbo3.5，拟合不同 prompt 对验证结果的影响

大纲

进度

- 通过 Charles 抓包，初步分析了校园集市平台的 http 包交互逻辑，通过无头浏览区构建了相应的 header，抓取了简单的数据
- 对于 SVM，朴素贝叶斯等构建了相应的实现方式

大纲

难点

- 集市平台采用 Vue 架构，html 骨架和层叠样式表以及 javascript 的结构不熟悉，需要花费一定的时间解析
- 对于 attention 机制的应用难度较大，需系统性的研读相关论文

项目结构 I

```
Sentiment Analysis

├── LICENSE
├── README.md
├──requirements.txt
├── main.py
├── dataset
├── src
│   └── __init__.py
│   └──data_processing
│   └──sentiment_dictionary
│   └──machine_learning
│   └──deep_learning
│   └──verification
└── reoprt
```

图: 文件组织结构

开源地址 I

https://github.com/anglee2002/SentimentAnalysis