汽车行业用户观点主题及情感识别

——数据科学导论期末报告

何秦兴 PB16110299

吴雨菲 PB15020525

张劲暾 PB16111485

2018 **年** 12 **月** 5 **日**

目录

1 题目描述

1.1 题目背景

随着政府对新能源汽车的大力扶植以及智能联网汽车兴起都预示着未来几年汽车行业的多元化发展及转变。汽车厂商需要了解自身产品是否能够满足消费者的需求，但传统的调研手段因为样本量小、效率低等缺陷已经无法满足当前快速发展的市场环境。因此，汽车厂商需要一种快速、准确的方式来了解消费者需求。

1.2 题目任务

本赛题提供一部分网络中公开的用户对汽车的相关内容文本数据作为训练集，训练集数据已由人工进行分类并进行标记，参赛队伍需要对文本内容中的讨论主题和情感信息来分析评论用户对所讨论主题的偏好。讨论主题可以从文本中匹配，也可能需要根据上下文提炼。

1.3 数据说明

训练集数据中主题被分为 10 类，包括：动力、价格、内饰、配置、安全性、外观、操控、油耗、空间、舒适性。

情感分为 3 类，分别用数字 0、1、-1 表示中立、正向、负向。

content_id 与 content 一一对应，但同一条 content 中可能会包含多个主题，此时出现多条记录标注不同的主题及情感，因此在整个训练集中 content_id 存在重复值。

张劲暾:

5.2 分工说明

吴雨菲:

何秦兴:

张劲暾:

6 参考文献