**基于推特数据爬取、清洗、情感分析的研究报告**

**摘要：**随着社交网络整体的普及程度不断加深，推特作为国外最主流的社交媒体越来越成为公众表达意见、传递情绪的重要渠道之一，基于此产生的呈现在公共视野中的各种观点、表述和词语也都有不同的情感取向和价值维度。对此类数据进行抓取并清洗，以此为内容的情感分析是本研究的主要方面。

1. **引言**

作为计算传播学科的重要组成部分和基础内容，对数据网络的爬取、清洗和情感分析有着重要的价值和现实意义。爬虫就是请求网站并提取数据的自动化程序，其中请求，提取，自动化是爬虫的关键。而数据清洗（data cleaning）则是数据分析的重要步骤，其主要目标是将混杂的数据清洗为可以被直接分析的数据，一般需要将数据转化为数据框（data frame）的样式，通常包括使用Pandas进行数据清洗、清洗错误行和列、提取所要分析的内容的基本流程。情感分析是基于清洗过后的数据进行的研究，是对爬取和清洗的进一步深化。

本研究希望对推特这一社交媒体平台上的内容、评论等诸多方面进行以上三个步骤的操作，在分析数据的过程中总计现象，并得出最终的研究结果。

1. **研究思路**

研究思路如下：

**（一） 数据来源**

选取2019.1.22的推特数据作为主要的爬取对象，网址为http://www.twiview.com。

**（二）数据爬取（crawl）**

考虑了3种获取数据的方式。

（1）通过 课程学习的bs4库爬取数据；

（2）通过八爪鱼、后羿等可视化采集器；

（3）通过 selenium 库模拟浏览器爬取数据。

经过实验发现，八爪鱼等采集器针对 推特 上这种通过动态加载的页面的处理不是很好，选择使用 selenium 库来进行数据爬取爬取速度较慢；最终仍然选择bs4库爬取数据。

在数据爬取部分，主要解决了两个问题：1）推特网站存在反爬虫的机制，大量异常的访问请求会导致IP地址被封而无法进行爬取；2）推特网站采用了动态加载而非翻页的内容更新方法，常常会导致爬取到同一页的重复内容。针对这两个问题，首先采用keys(10个代理)防止因为访问量异常导致IP地址被封，随机从10个代理中挑选进行访问；由于推特网站使用动态加载技术而非翻页加载的技术，同时模拟动态加载的网页进行访问，以访问其他页的内容。为了加快访问速度，添加每15s添加线程进行一次新的访问。

本次爬取的数据主要包括网站上的文字信息内容以及各类评论，因为网站对IP进行访问限制以及时间所限，数据爬取有数量上的限制。所以爬取的数据总量大概是条。

**（三）数据清洗（clear\_data）**

由于推特网站上存在大量重复的文本，因此需要对推特数据进行清洗，同时只保留保留中文或者英文的评论，以保证能够被学习处理。

将爬取的数据分类整理，包括将视频的简要信息放到一个xls中，保留中英文评论分开。过滤中英文评论，去除掉数字、标点符号及表情等，去除无效的评论，例如只有标点符号的评论。

具体地，通过正则表达式，分别提取出中文加数字和英文加数字，分别进行清洗处理。

然后进行手动加标签（0/1），0代表消极情感，1代表积极情感，作为验证集。

**（四）情感分析 (text\_classifier)**

情感分析可以看作一个正向/负向的二分类问题，可以使用分类模型判断文本的情感。

情感分类模型常见的有RCNN，RNN，和引入注意力机制的 网络几种；本次实验选择了引入注意力机制的RNN网络，通过引入attention的概念改变不同关键词的权重，可以提高情感分类模型的准确度。

由于提取到的数据量少，中英文文本采用推特中公开的语料进行训练，挑选出部分语料作为测试集进行测试（在附件中en是英文公开语料，ch是中文语料），观测其准确程度。由于中文语料较少，因此中文测试集也是从公开数据集中挑选并补充推特中数据获得。分别对中英文进行处理。

comment中包含词典分类的功能：词典是文本挖掘最核心的部分，原始词典按照习惯将词汇分为：情感词（积极评价词、积极情感词、消极评价词、消极情感词）、程度词（从最重的 most程度依次降低到 least 程度，共 6 个等级）和否定词三类。 可以提取相关的关键词展示出相关的情感。

**三、发现**

通过以上情感分析模型，可以总结出推特上每天都有各类主题的内容展示和各类不同方向的情绪或者情感表达，内容角度各不相同。除此之外，中英文评论也一定程度上都会关照当下的社会现实，且中外受众在一些具有普适性的问题上呈现出相似性的特征，但也有所差别。

**四、结论**

推特作为国外最主流的社交媒体网站之一，每天打开、浏览、评论的用户和受众始终保持在较高水平，作为社会现实的风向标以及社情民意的显示器，推特上的相关内容以及与此有关的各项评论数据都是本研究的主要内容，中英内容以及基于此所反映出的国别差异也可见一斑。

本实验从不同的维度，包括基本的数据挖掘、清洗、评论主题词提取以及情感分析等，用数据说话，用数据用对比的方式从多个维度对推特这一社交媒体内容情况进行研究，由小见大，窥探到更深层次的语义差别。