**项目阶段报告**

田 也 3120181028

商瑞红 3220180732

王若琳 3120181040

王孟岚 3120181037

**数据获取及预处理**

使用爬虫技术。在选择爬取内容时，考虑到携程、淘宝、天猫以及京东的商品评论无法直接获取到，都使用了ajax加密技术，需要从另外的页面中读取评论数据。虽然经过对网络流量进行监控，我们获取到评论的真实存储地址，并尝试进行爬取，但他们使用了第二重防爬虫策略：动态cookie，每次向评论存储地址发出请求时，浏览器都会传送一个cookie，服务器进行检测，通过检测后才会返回数据，否则只能访问到拒绝访问的json返回值。同时由于cookie是动态变化的，无法进行预测，爬虫程序的编写难度较大。因此通过小组讨论，我们决定改为对网易云音乐评论进行单句分析。具体的实现方式是对前端输入的网易云音乐URL进行解析，若解析不成功则进入错误提示页面，解析成功则生成对应的短评页面url。网易云音乐的短评数量多则上万条，为了减轻系统压力，网易云爬虫爬取最新的前200条短评，并生成短评列表返回。

**实验环境**

(1) 客户端操作系统：Windows/Mac OS/Linux

(2) 浏览器：chrome v50.0以上/Firefox v42.0以上

(3) 系统服务器：python 2.7.10+

(4) 使用库资源： django 1.11

BeautifulSoup4 4.6.0

Jieba 0.39

Requests 2.18.4

(5) 程度副词词典，来自互联网

**实验模型及流程**

本系统使用python语言实现，以网站形式呈现，网站使用django架构搭建。系统的整体架构具体分为评论爬取模块、分词模块、分析模块、极性分值模块、前端展示模块及双树模块。其中，评论爬取模块为情感分析提供原始资源，每条评论以字符串形式存在，多条评论组合成一个列表；分词模块将每一条评论拆分为单个词，并标注词性；针对单条评论和多条评论，分别使用单项评论分析模块和整组评论分析模块；频繁模式树和最小生成树为评论分析提供支持；极性、分值模块计算、存储每一条评论所得到的分值，并判断整组评论的极性为积极或是消极；前端展示模块接受服务器传送的数据，同时将处理得到的用户情感分析结果以可视化的方式展示在界面上。

**待完成工作**

在对算法进行调研之后，决定选择基于规则的FP树算法来实现单句的情感分析功能，这个算法能保持数据完整性:它不会打破任何事物数据中的长模式，而且为频繁模式的挖掘保留了完整的信息。也有紧凑性的优点，减少了不相关的信息，非频繁的项会被删除，按频率递减排列，相似的更频繁的项更容易在树中被共享。但由于对此算法不熟悉，所以集成时遇到了一些问题，我们目前也正在阅读论文。另外系统整体框架已经搭建好，但编写过程中关于前端实现的一些细节问题还需要进一步完善。