总结

# 一、作业要求

## 毕业项目：构建一个舆情分析平台

**项目背景：**某公司计划新上线一款苏打水饮料，为了了解用户对苏打水的接受程度，需要抓取“什么值得买”(<https://www.smzdm.com/fenlei/qipaoshui/>) 网站中气泡水种类前 10 的产品的用户评论，通过对用户评论的正向、负向评价了解排名前 10 的气泡水产品的用户接受程度。

## 具体需求：

1. 正确使用 Scrapy 框架或 Selenium 获取评论，如果评论有多页，需实现自动翻页功能，将原始评论结果存入 MySQL 数据库，并使用定时任务每天定期更新。
2. 对评论数据进行清洗（可借助 Pandas 库），并进行语义情感分析，将分析结果存入数据库。
3. 使用 Django 集成在线图表对采集数、舆情进行展示，需包括该产品正、负评价比例，以及评价内容等。
4. 数据展示支持按时间筛选和按关键词筛选功能（参考百度情感分析后台 [https://ai.baidu.com/tech/nlp\_apply/sentiment\_classify）。](https://ai.baidu.com/tech/nlp_apply/sentiment_classify%EF%BC%89%E3%80%82)

## 评分标准：（实现相应功能，每项 +10 分，部分实现 +5 分）

1. 正确使用 Scrapy 框架获取评论，如果评论有多页，需实现自动翻页功能。
2. 评论内容能够正确存储到 MySQL 数据库中，不因表结构不合理出现数据截断情况。
3. 数据清洗后，再次存储的数据不应出现缺失值。
4. Django 能够正确运行，并展示采集到的数据，数据不应该有乱码、缺失等问题。
5. 在 Django 上采用图表方式展示数据分类情况。
6. 舆情分析的结果存入到 MySQL 数据库中。
7. 在 Django 上采用图表方式展示舆情分析的结果。
8. 可以在 Web 界面根据关键字或关键词进行搜索，并能够在页面展示正确的搜索结果。
9. 支持按照时间（录入时间或评论时间）进行搜索，并能够在页面展示正确的搜索结果。
10. 符合 PEP8 代码规范，函数、模块之间的调用高内聚低耦合，具有良好的扩展性和可读性。

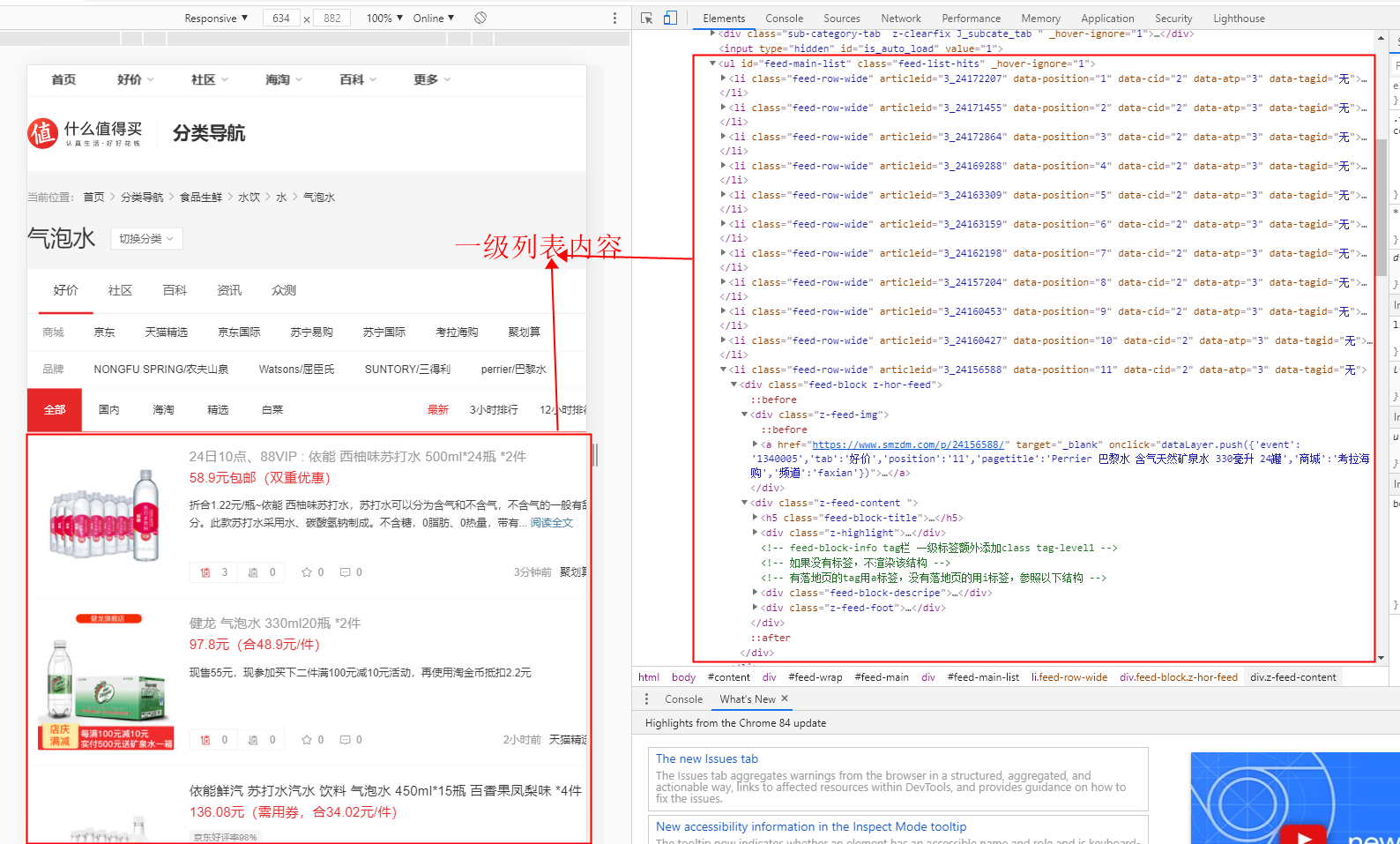
# 二、需求分析

### 爬取数据子项目需求分析

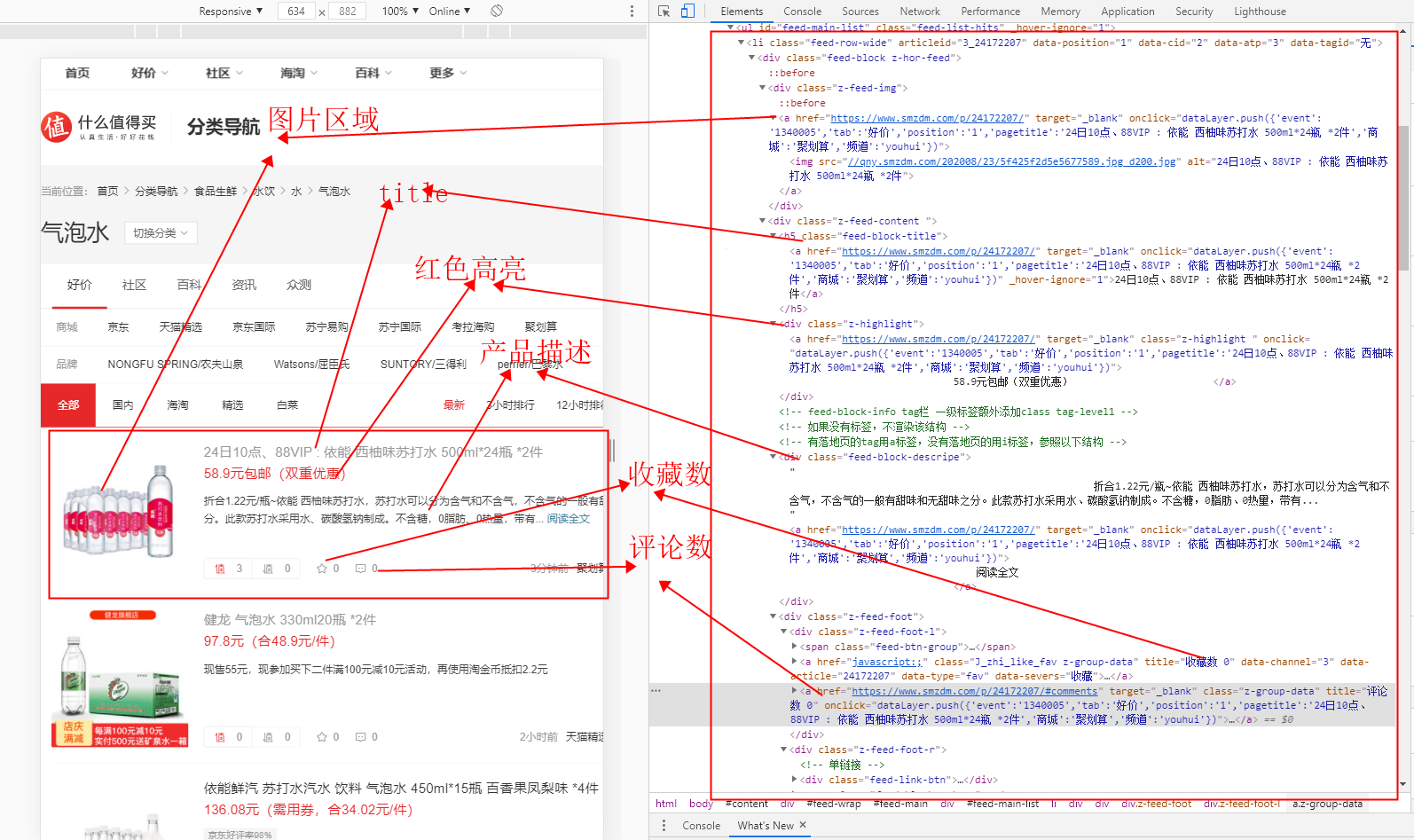
原始数据特征分析

#### 主页面一级列表容器

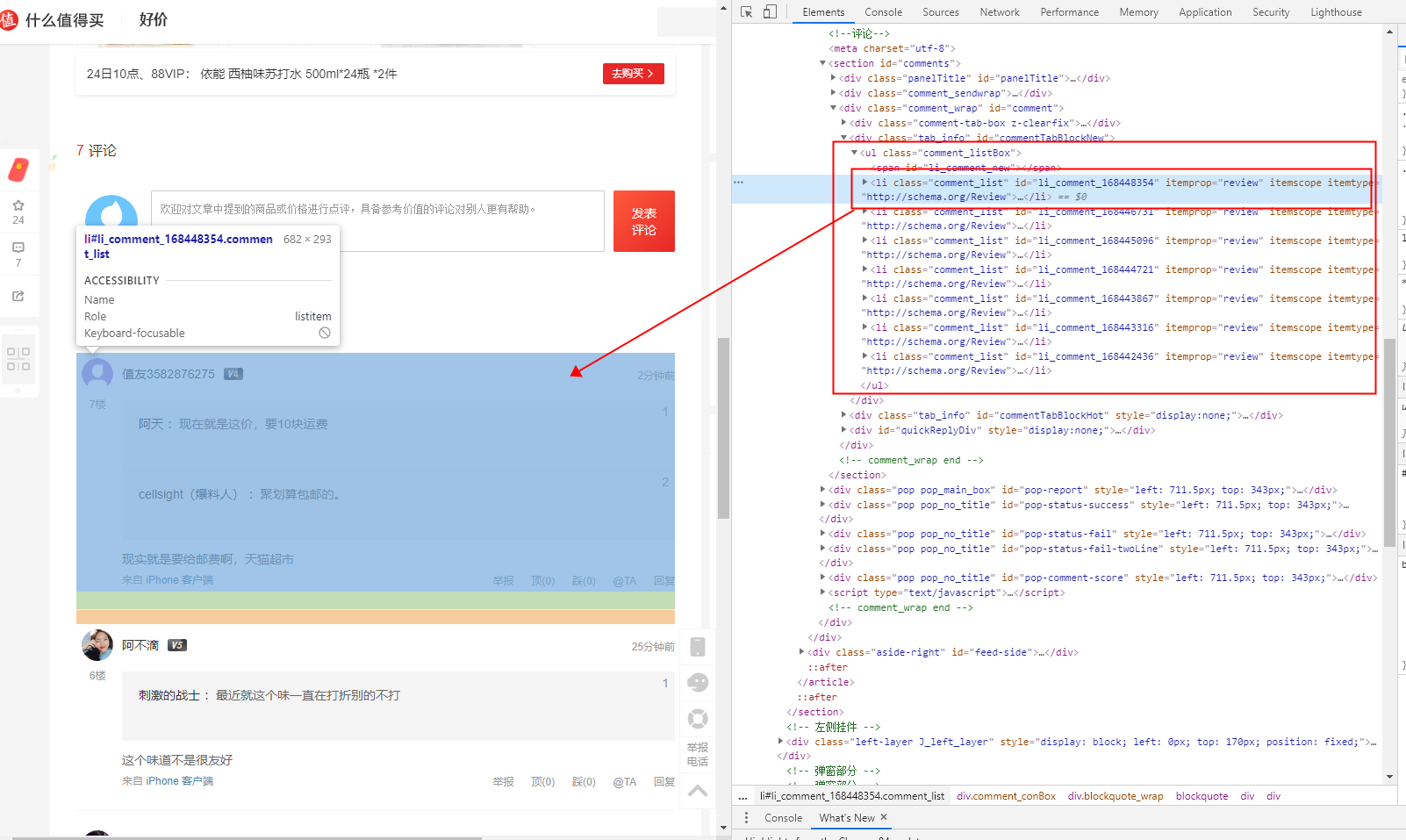
ul的id为feed-main-list



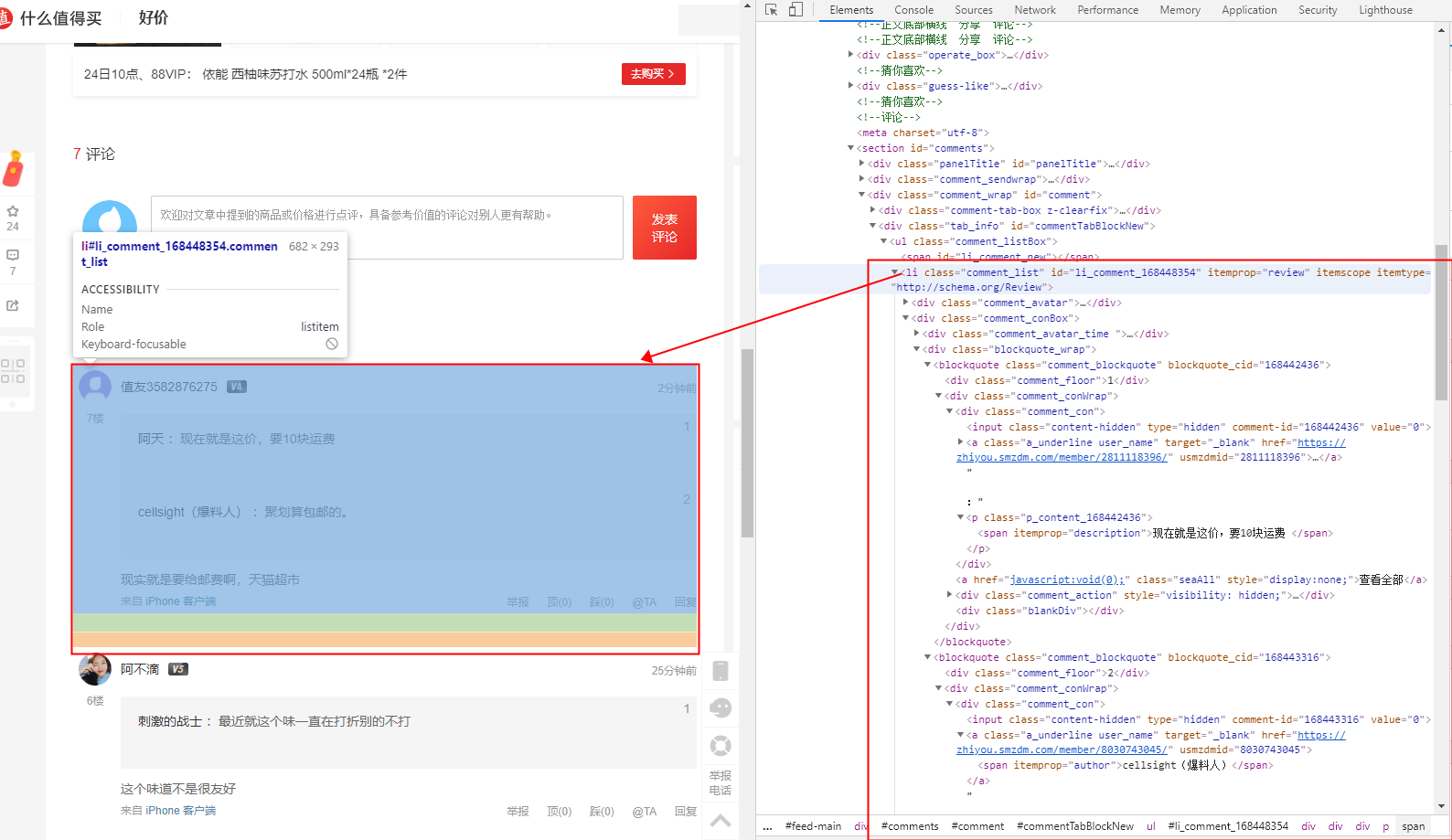
#### 主页面二级容器

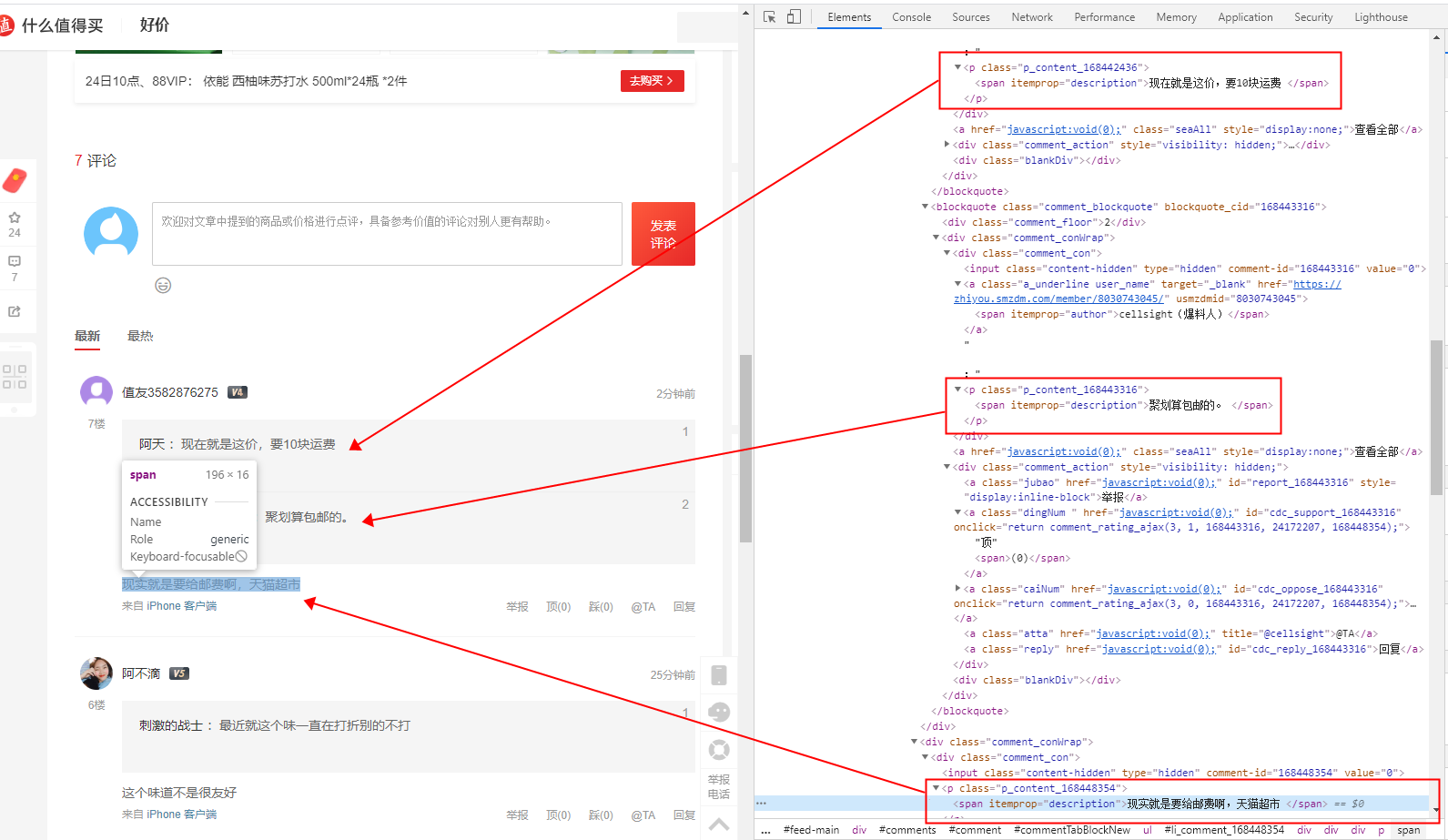


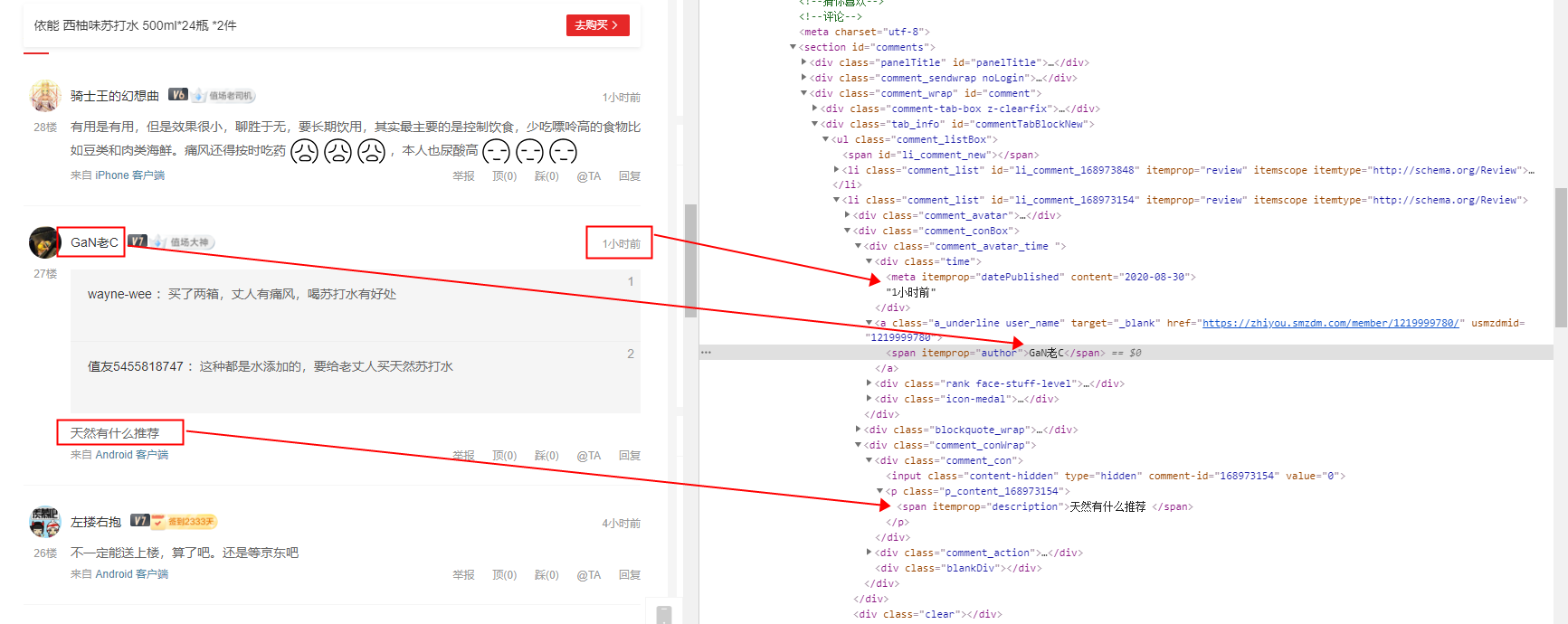
#### 详情页面评论一级容器



#### 详情页面评论二级容器







# 三、最终效果图

