

**《数据挖掘实验项目》**

**分析报告**

|  |  |
| --- | --- |
| 学 院： | 信息工程学院 |
| 专 业： | 信息管理与信息系统 |
| 年 级： | 2018级 |
| 组 号： | 五组 |

信息工程学院制

2021年01月

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| 年级 | 2018级 | | 专业 | | 信息管理与信息系统 | 班级 | 二班 |
| 组号 | 五组 | | 队名 | | 对对队 | | |
| 队呼 | 对对队 做的全会 蒙的全对 | | | | | | |
| 主题 | Steam 2077评论挖掘与分析 | | | | | | |
| 小 组 成 员 | | | | | | | |
| 姓名 | | 学号 | | 组内分工 | | | |
| 肖芙蓉 | | 201810520224 | | 爬取数据、数据处理、撰写报告 | | | |
| 栾启彤 | | 201810520204 | | 编写代码并调试运行、数据可视化处理、 汇报PPT | | | |
| 田锦 | | 201810520226 | | 爬取数据、数据处理、撰写报告 | | | |
| 郝秀丽 | | 201810520212 | | 爬取数据、数据处理、制作PPT | | | |
| 刘晋妤 | | 201810520130 | | 获取各类游戏资料、数据处理、日志的编写及汇总 | | | |
| 成绩评定 | |  | | | | | |
| 组长签字 | |  | | | | | |
| 教师签字 | |  | | | | | |

1. **问题背景**

《赛博朋克 2077》是一款开放世界动作冒险游戏，故事发生在夜之城。这是一座五光十色的大都会，权力更迭和身体改造是不变的主题。

但是《赛博朋克2077》售价高达298元，相当于普通大学生一周的生活费。那《赛博朋克2077》是否值得学生拿出一周生活费去购买一个游戏，这值得仔细探讨，普通网友了解游戏是否值得购买的途径只有看游戏下面已经购买游戏的玩家的评论和视频平台上的试玩视频。

但目前《赛博朋克2077》在Steam上的评价由预购时期的“特别好评”变成了“多半好评”，目前游戏的中文语言评价为“褒贬不一”。各种技术原因是造成口碑下跌的主要原因，很多玩家反映优化太差，有很多bug以及无法进入游戏等等。游戏评论鱼龙混杂，试玩视频太过片面，在此情况背景下，本项目诞生了。

《赛博朋克》为什么会迅速风靡全网？赛博朋克2077真的好玩吗？已购买玩家对未购买玩家有什么建议？现在的玩家对2077的态度是什么样的？这一系列问题，每位玩家都有自己的不同看法，那如何正确提取其中有用信息，合理采取判断这些评价，这是我们选取此题目的目的。

通过对《赛博朋克2077》评论进行分析挖掘，对游戏评论进行爬取清洗筛选，根据评论进行语义分析玩家对游戏的态度，根据分析结论生成可视化图表、词云、柱状图、饼图，根据玩家评价做出游戏推荐购买度，为后续玩家提供参考基础。

1. **提出问题**

《赛博朋克2077》评价褒贬不一，该如何合理客观看待这些评价，从中选取有价值信息？这些需要数据挖掘技术，对游戏评论进行提取挖掘分析，做出合理判断。生成结果之后，可以为以后玩家提供参考，也可为游戏提升优化提出合理建议。

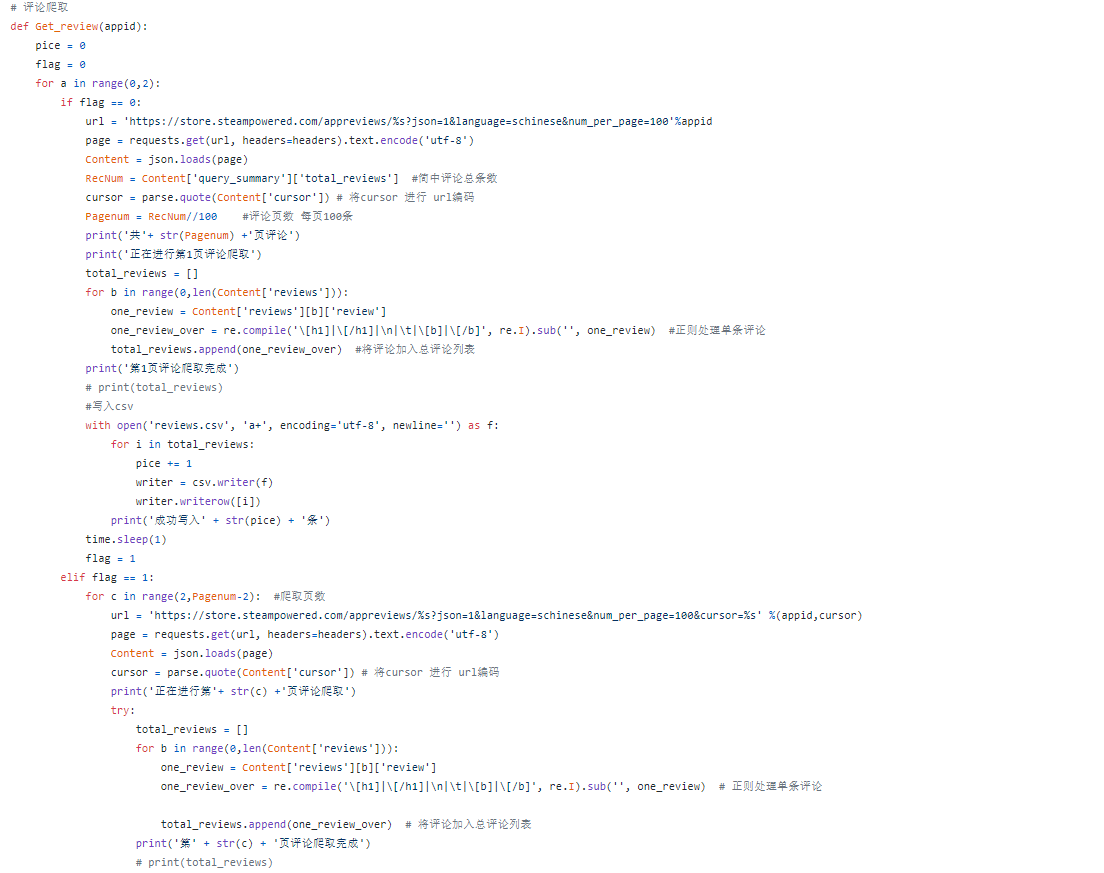
1. **分析问题**
2. 游戏的评论从哪里来？
3. 游戏的评论是否完整？
4. 该如何选用数据分析算法?
5. 训练集的语料库从哪来？
6. 模型评估模型准确率怎么算？
7. 用哪种图表展示数据更加直观？
8. **数据获取方法及过程（附主要代码）**
9. 训练集的获取：

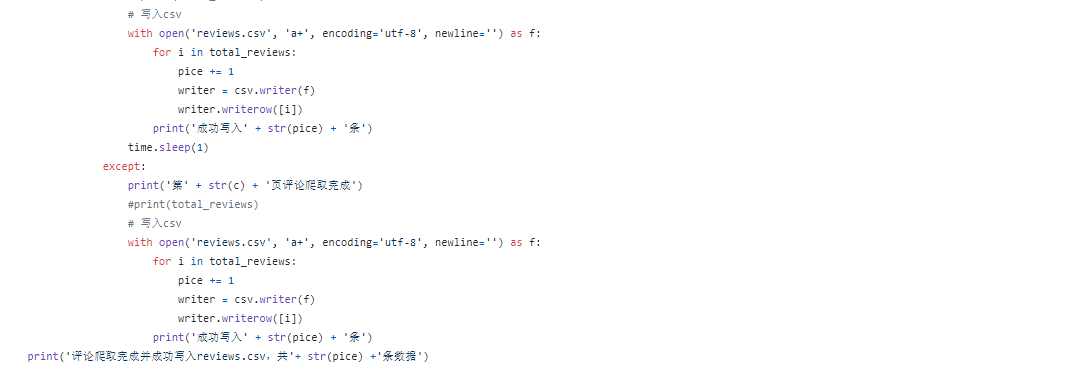




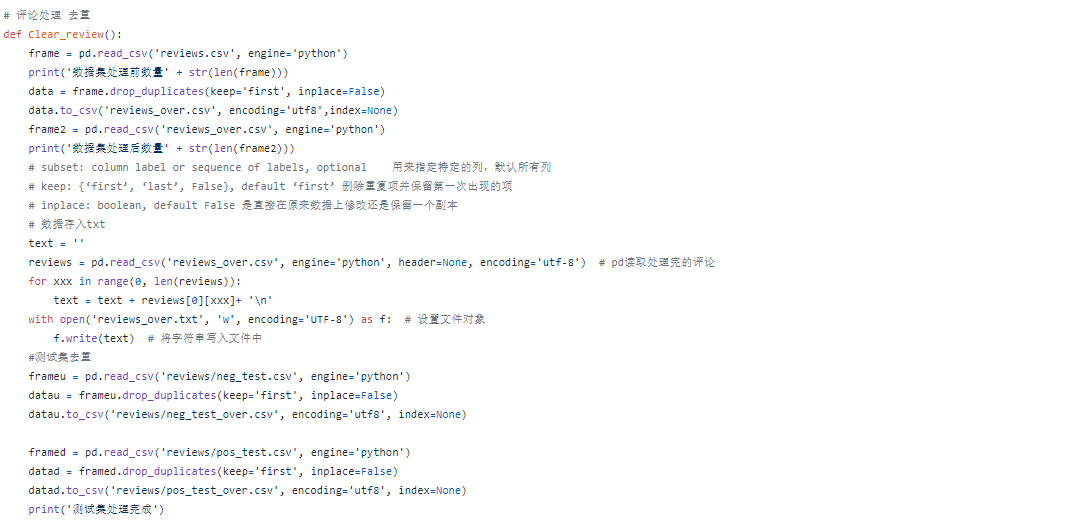


1. 数据集的获取：

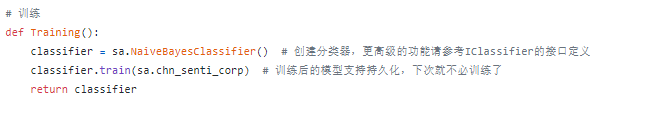




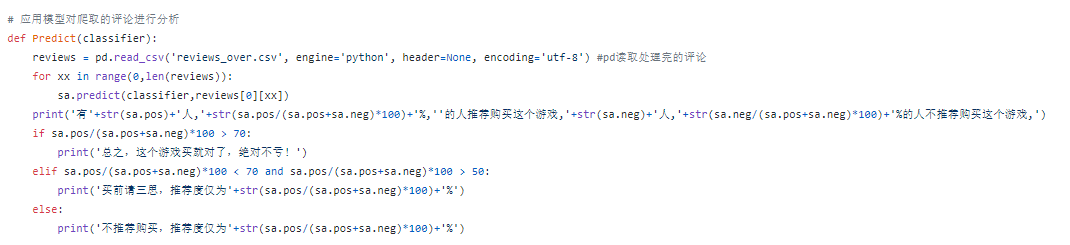
1. **数据挖掘过程（分析数据、数据清洗过程）**
2. 数据处理，去重：



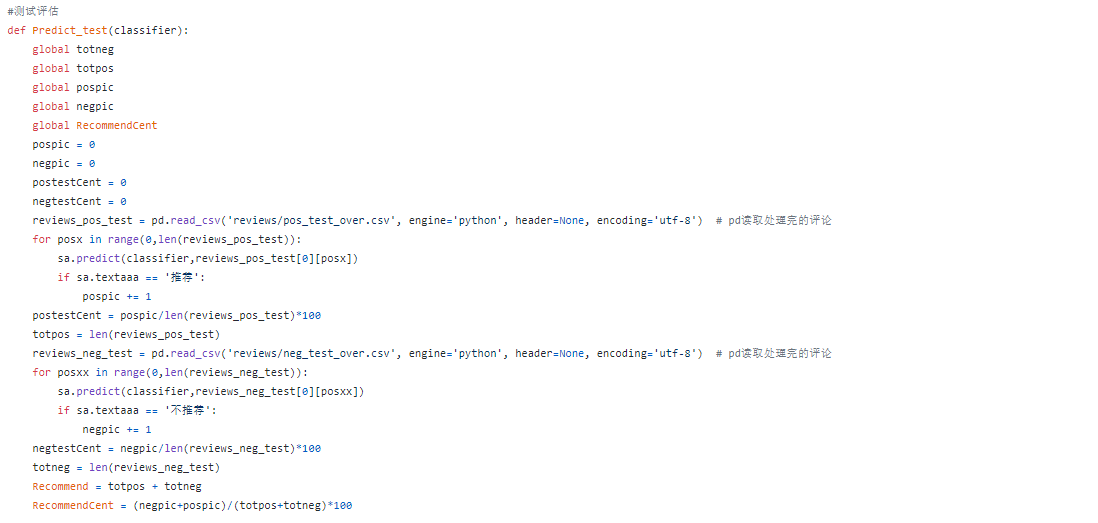
1. **模型训练过程：**



1. **模型应用过程：**



1. **测试评估：**



**九、数据可视化**







1. **结论**



图1 词云图

1. 通过图1可以知道，《赛博朋克2077》游戏评论中，关键词BUG最多，剧情次之，其中还包含垃圾，修复，差评，差等负面词语。但也包含好评等正面词语。但负面词语居多。说明《赛博朋克2077》游戏暂时不推荐购买。

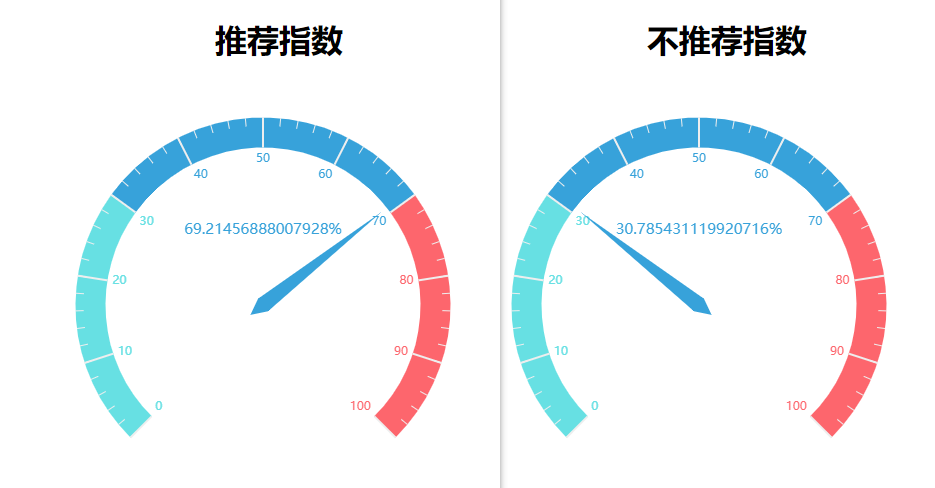
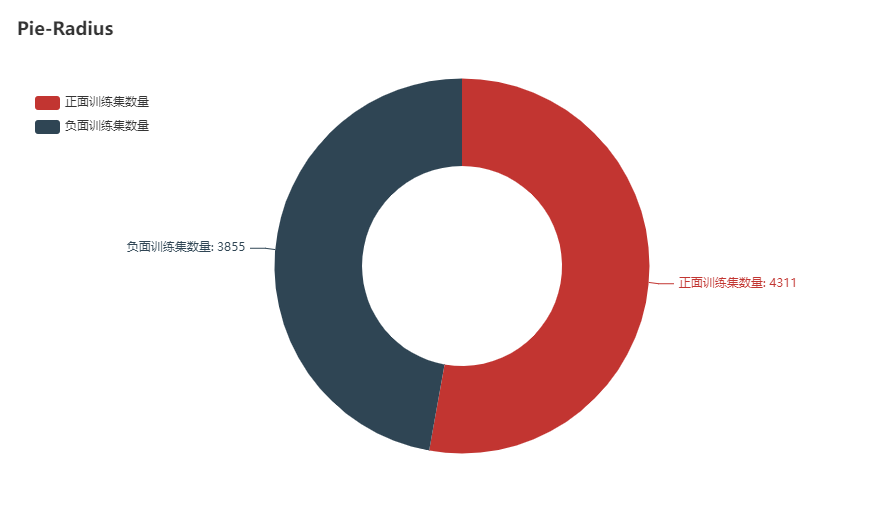


图2 推荐指数仪表盘

1. 由图2可得，本项目通过训练模型对爬取的评论进行分析得出，《赛博朋克2077》游戏的推荐指数仅为69.2%，为需要仔细斟酌购买。结论与结论1一致。

图3 训练集数量

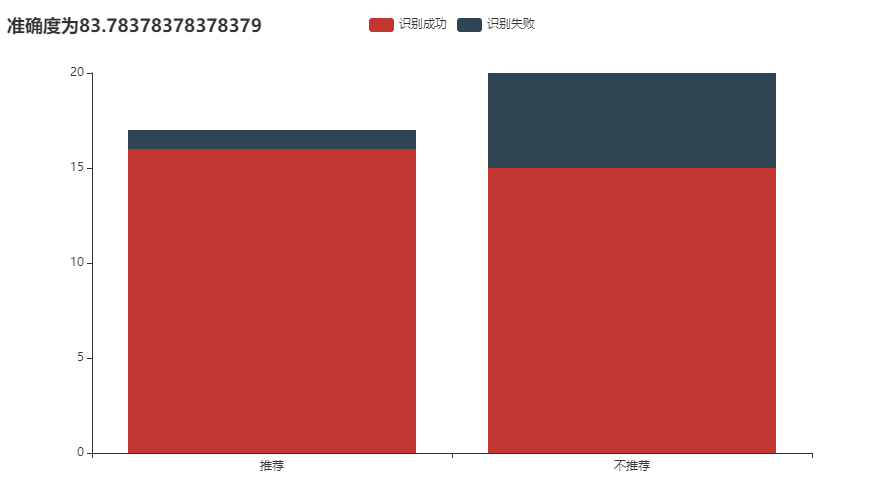


图4 模型训练准确度

1. 从图3、图4可以得出，本项目语料库正面训练集数量为4311，负面训练集数量为3855。模型识别准确度为83.78%，为中等偏上水平。
2. **收获**

通过这次项目，我们小组分工明确积极完成每个人的任务。在实验过程中，遇到问题，大家相互讨论，搜集资料，解决问题。

在项目完成过程中，我们先从数据爬取下手，利用了上课学到的爬虫技术，将游戏评论爬取下来，并分为好评和差评。为了使得数据更加完善，同时学习了HanLP算法，使得本小组的项目功能更加完善，结构更加清晰。接着，通过HanLP算法统计了游戏中好评与差评的词频，并制作了词云图，使得数据一目了然。将爬取到的数据，进行建模，

完成了这次任务，我们深刻认识到，数据挖掘是来源于生活，服务于生活，与我们日常贴近。数据挖掘有很多知识和算法支持，我们要不断学习，充实完善自己。将课上知识与课外知识结合，更好的掌握。在以后学习中，不断应用补充完善。