总体设计

研究背景

数据的预处理

- 数据提取（使用正则表达式提取游戏时长中的小时数值）
- 去重（共去除1004条重复）
- 缺失填充（3条其他外语翻译为英语）
- 中英区分（使用正则表达式）
- 1. 评论含中文记为中文评论
- 2. 评论仅含英语记为英语评论
- 3. 评论纯英语但用户名称含中文额外记为国人英语评论
- 4. 中英双语评论保留中文部分

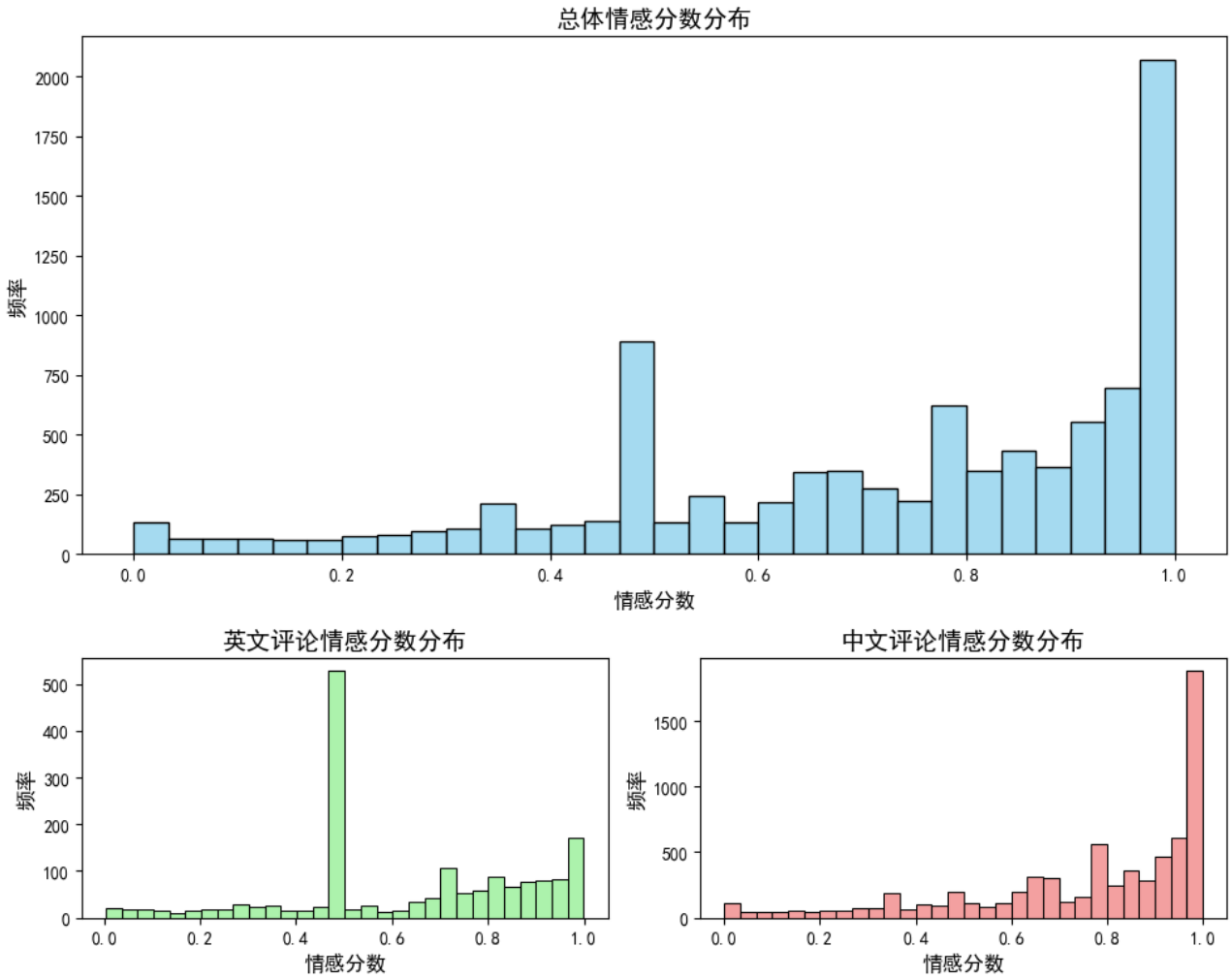
数据的初步统计

- 词频统计（词云）
- 游玩时长直方图
- 评论长度直方图
- 各语言评论数柱形图
- 中英用户的推荐数饼图

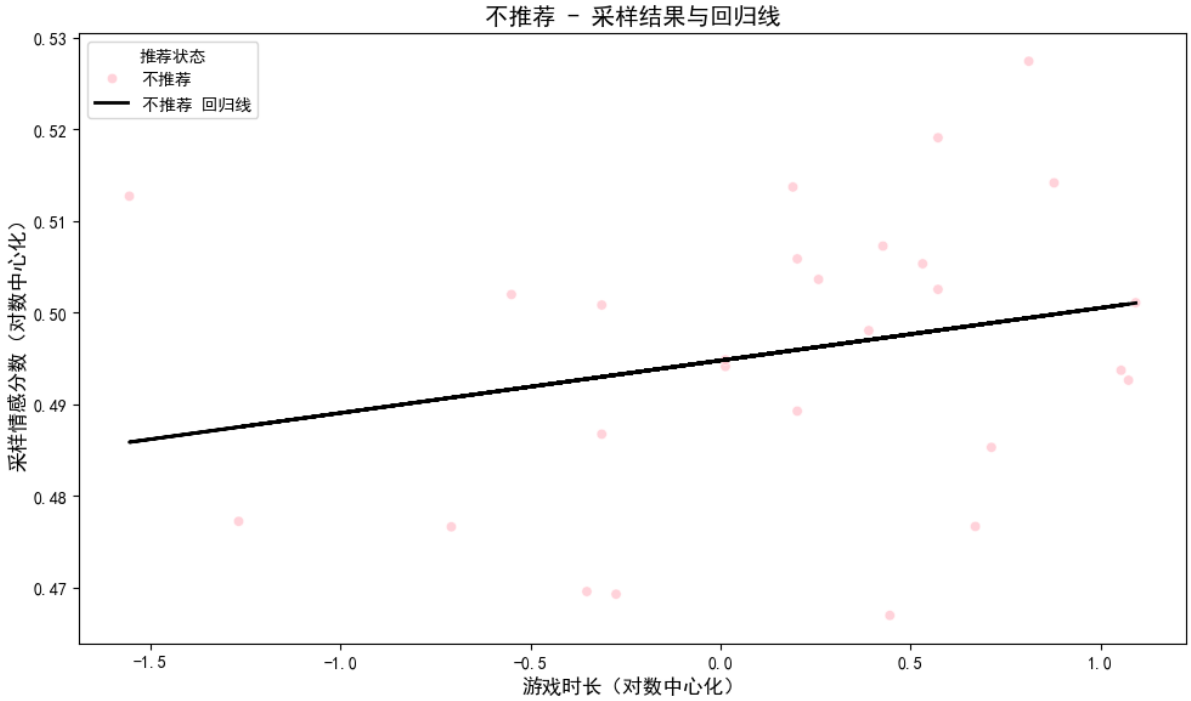
玩家态度分析

- NLTK情感分计算（简单介绍原理）

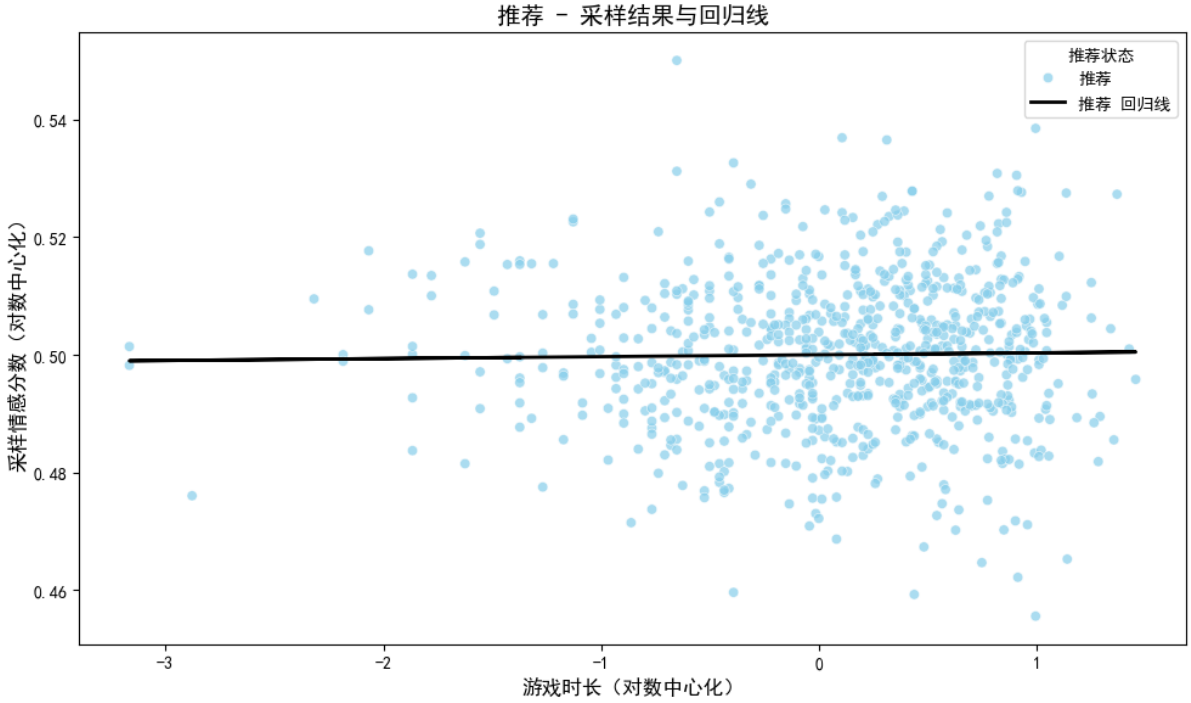
- 情感直方图



- 情感分对游玩时间做回归
  - 对数据的处理：
    1. 对情感分数和游玩时长都进行了对数中心化
    2.
  - 对不喜欢本游戏的玩家，初始印象所造成的情感分数截距较低，但评论态度会随着游玩时长的增加而有所好转，说明游戏本身质量过硬。（斜率: 0.0057, 截距: 0.4948）



- 对喜欢本游戏的玩家，其游玩时长对评论态度的线性相关性相对较弱，初始印象和评论主体使得截距出于0.5的理性无情绪状态，但也有缓慢上升迹象。（斜率: 0.0003, 截距: 0.5000）

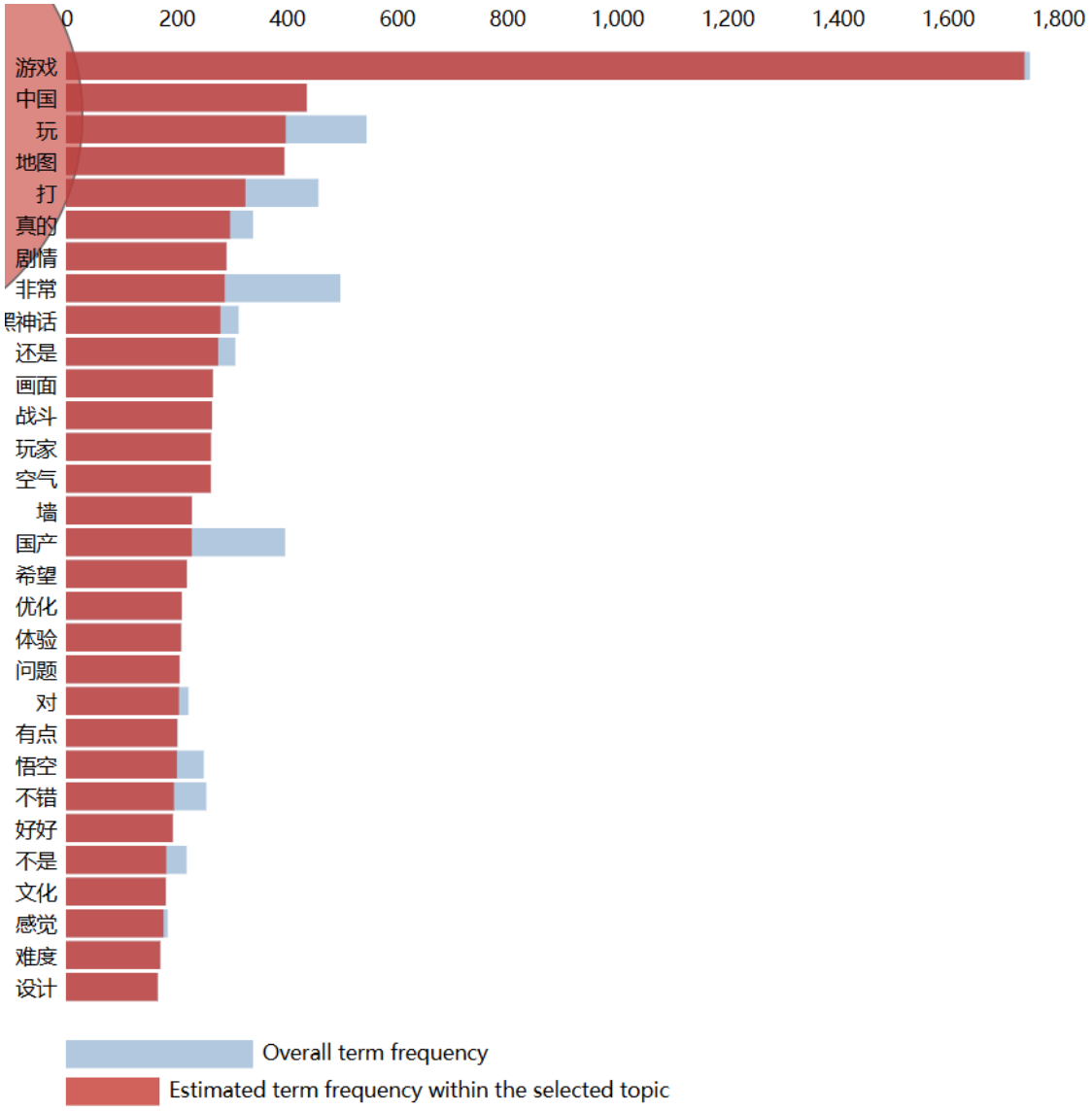


- 给出分析结论并对比中英评论

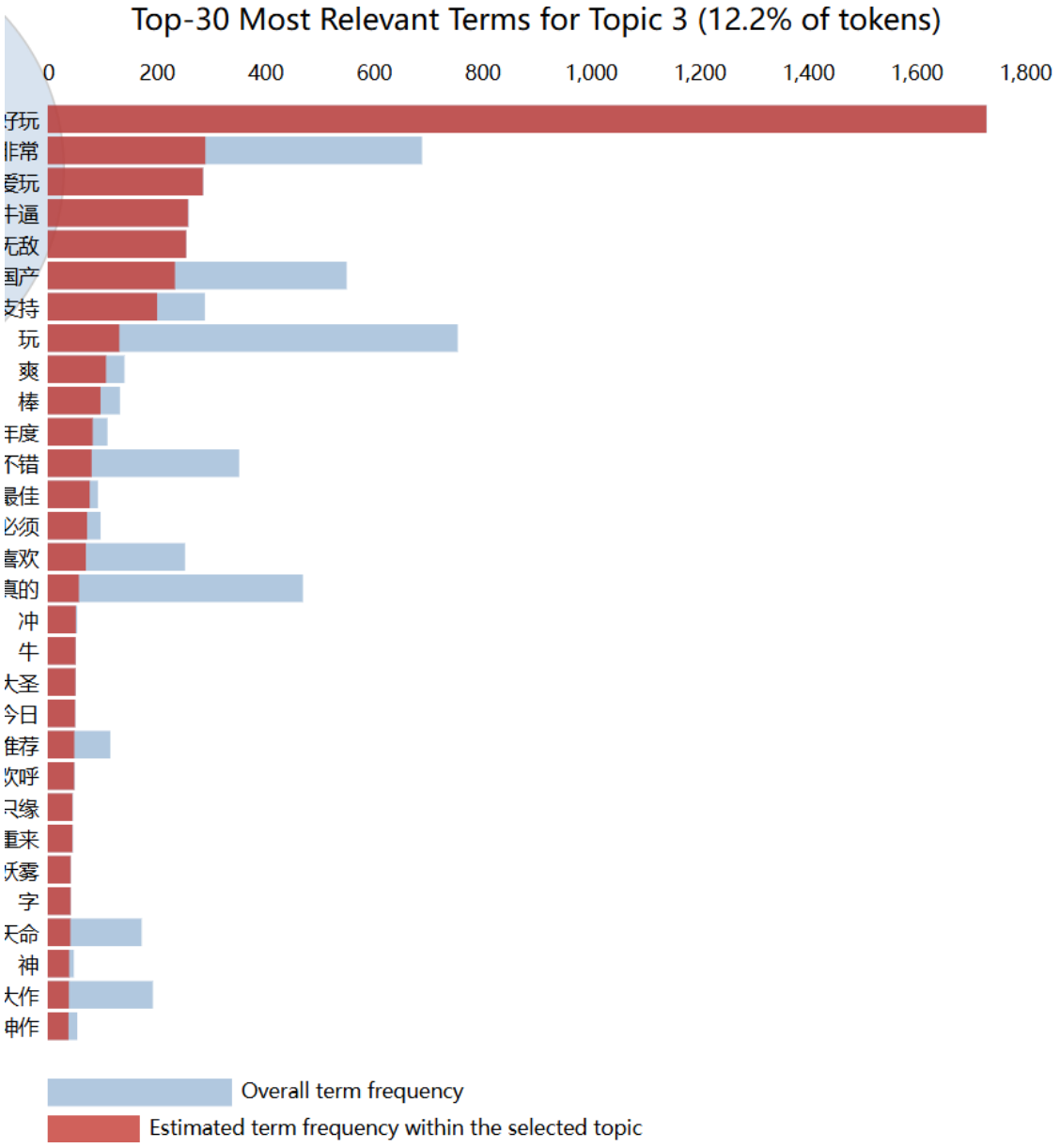
LDA主题分析（弃用此版，采用第一版）

- 分中英LDA建模
- 对不同主体内进行简单统计分析然后命名各个主题
  - **中文评论**
  - 推荐

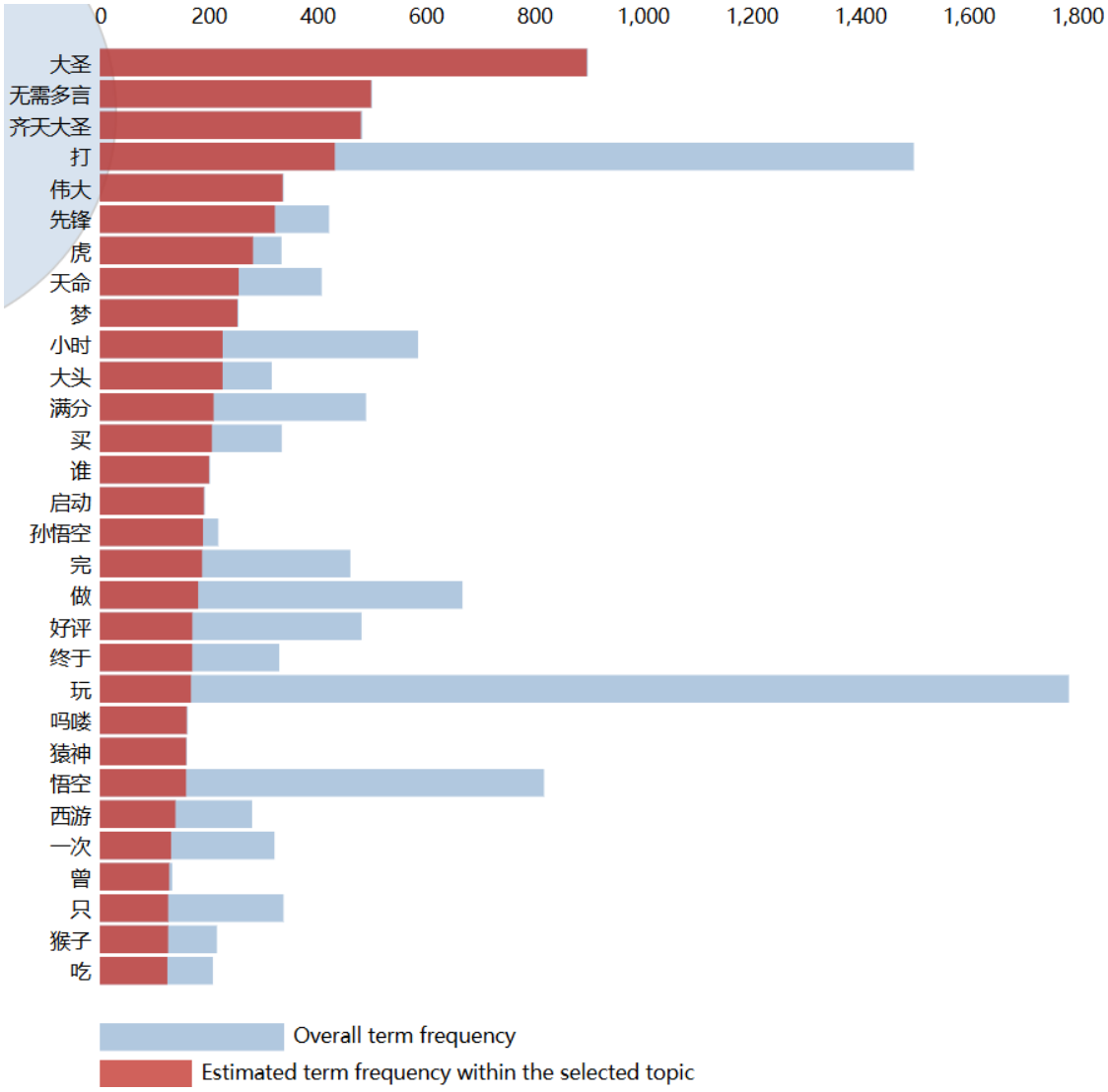
1. 大部分评论是“讨论派”（讨论游戏具体体验，也会讨论部分问题，占**67.6%**）



2. 少部分评论是体验党（对游戏体验给予更直白强烈的溢美之词，但没有过多具体内容的讨论，占**12.2%**）

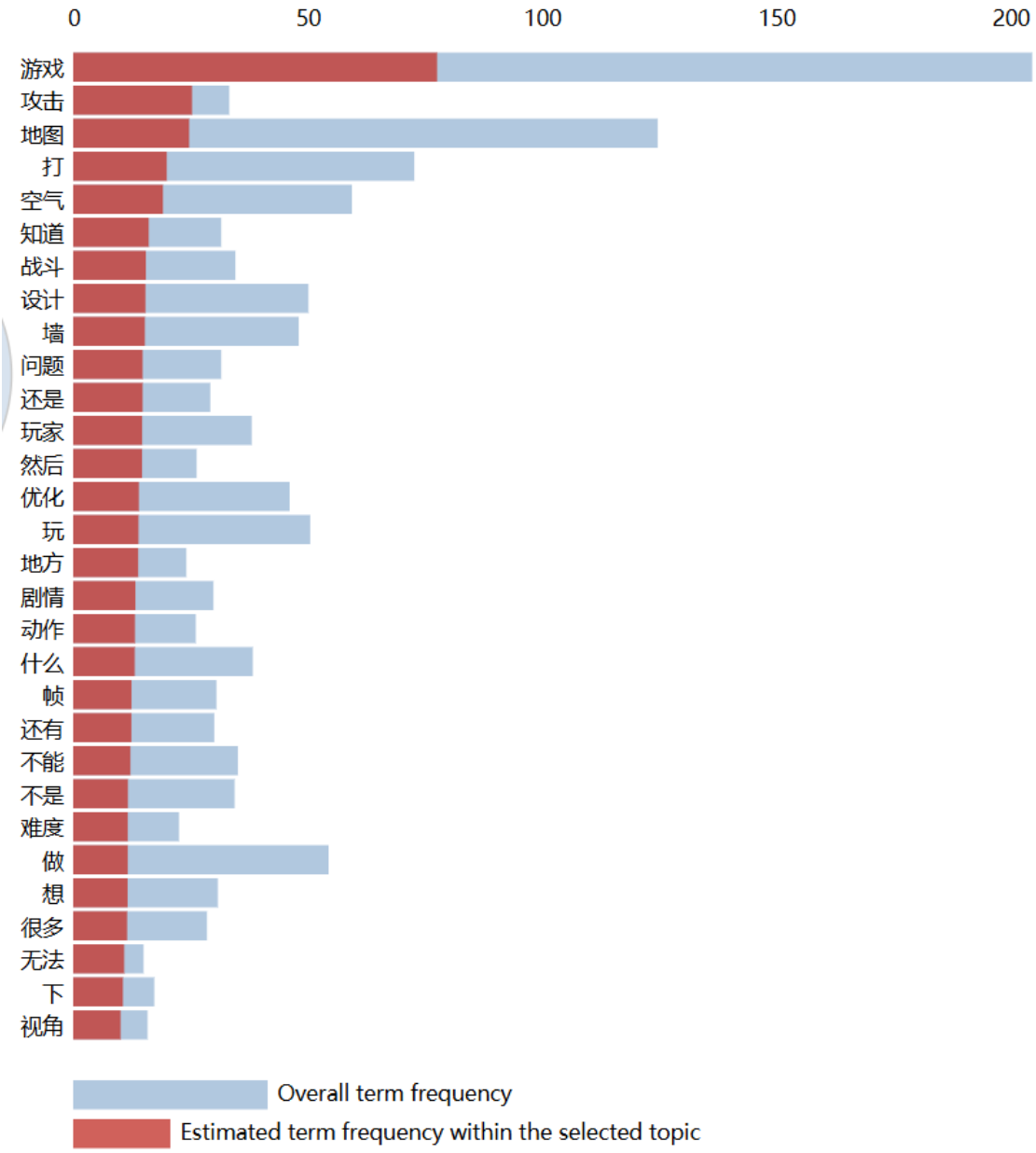

Top-30 Most Relevant Terms for Topic 3 (12.2% of tokens)
0
200
400
600
800
1,000
1,200
1,400
1,600
1,800
Overall term frequency
Estimated term frequency within the selected topic

3. 还有一类是“文化派”（譬如对“国产”或是“齐天大圣”提及较多，占**20.2%**）

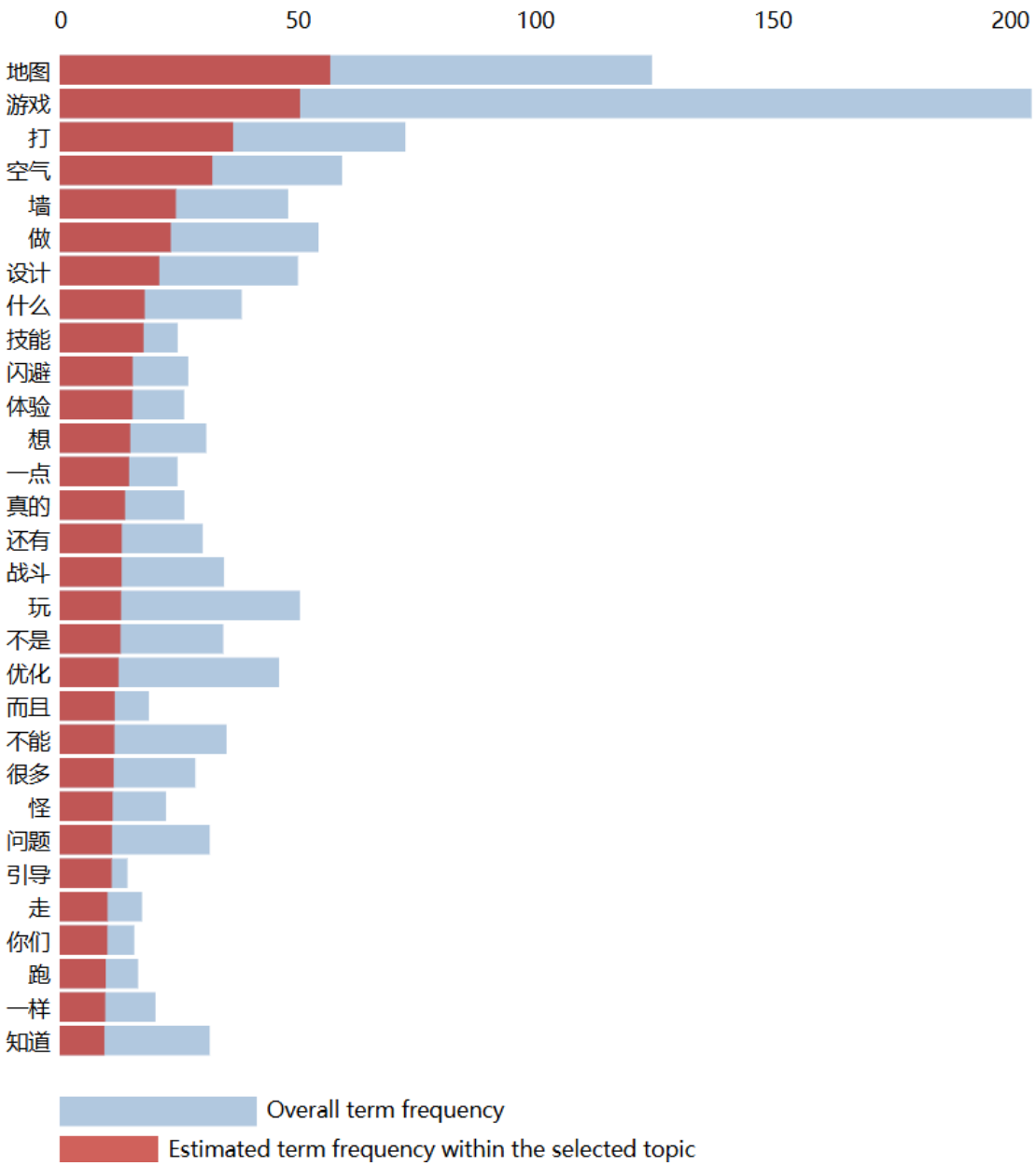


- 不推荐

1. “技术不足型”（针对游戏动作系统的吐槽以及一部分对向美术妥协的空气墙设计+对于较弱性能设备的优化不足问题的不满，占**35.5.5%**）



2. “内容不适口”（对游戏的没有地图的引导缺失、剧情、动画设计等技术经验领域的吐槽，占**36.6%**）



3. 以及混合了前两派理由但混合了大量谩骂的不礼貌评论（占比**20.5**%）

- **英语评论**

- 推荐

  1. 绝大多数评论是“讨论派”但讨论游戏问题的更少，更多的是对来自中国游戏“奇迹”的赞美和体验的优秀（国人英语评论虽然多是些简短的溢美之辞，但也多在此类，占**76.3%**）

  2. 再者是“香槟党”，不吝溢美之词且直呼游戏可以赢下今年“年度游戏”桂冠（占**5.8%**）

  3. 则是文化上的“好奇派”，由于黑神话悟空这款游戏产生了对中华文化和民间传说的好奇（**占比17.9%**）

- 不推荐

  1. 对战斗系统和视觉效果的“内容性吐槽”（从措辞来看大概率是`combat setting`之类的动作组合不流畅，结合专业媒体分析是预输入设计缺失引起的难度过大，占**54.6%**）

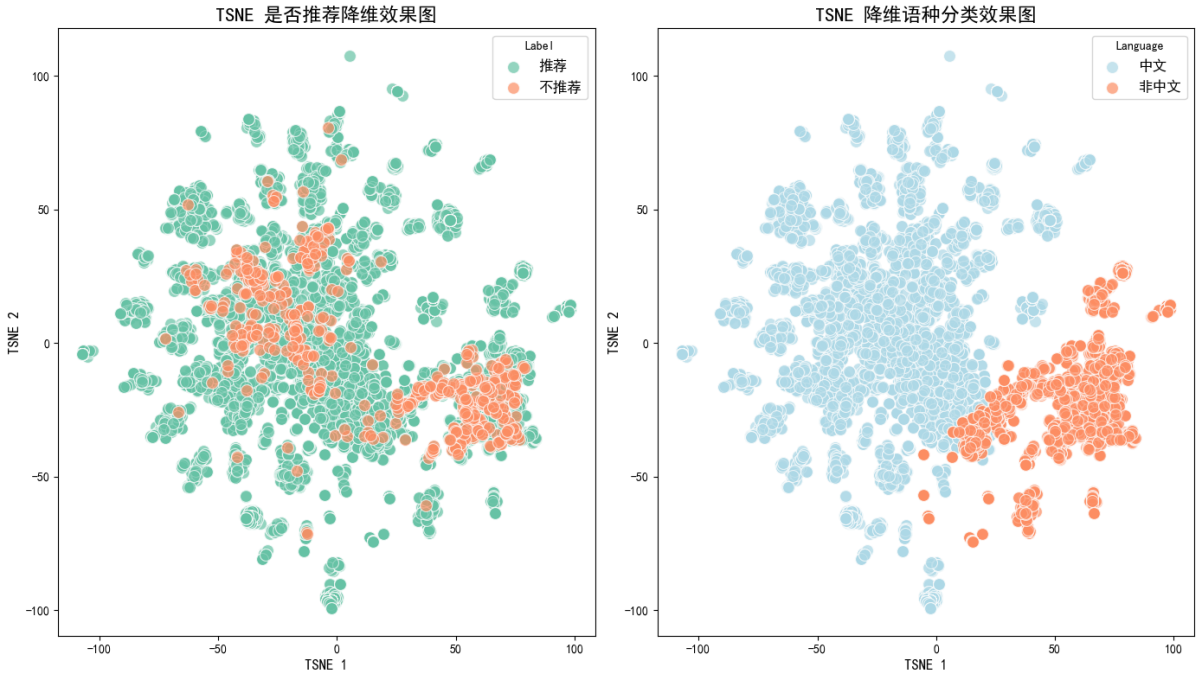
  2. 对游戏的性能和优化等“技术性问题”提出批评（譬如游戏崩溃，帧率低等，占**38.6%**）

3. 对于游戏中空气墙，叙事节奏、音效和严重性能问题等可以靠游戏开发经验积累得到改善的问题并没有被大肆讨论（占**6.8%**）

- 使得主题变为一个特征

对玩家群体内在关系的进一步探索

- 编码和降维流程简介
    1. 使用中英混合精校分词的`TF-IDF`方法编码出25692维的原始编码
    2. 使用`TruncatedSVD`方法将原始稀疏数据降维到50维
    3. 使用`t-SNE`方法降维到2维并执行分类可视化工作
- t-SNE
    1. t-SNE降维后的各语言样本分布图



    2. 分析：

1. **注意到经过分层降维后的文本数据在空间上主要呈现语种分离的空间分布，而各语种中“不推荐”的评论均分布在样本点相对靠近语种质心的位置，可通过计算样本特征散度构建新特征。**
2. 量化数值分析（类内散度的计算）

类别	推荐的散度	不推荐的散度	内部散度	不推荐点分散度
中文评论	5702.709	1530.442	5546.549	0.27592684476092066
飞中文评论	2569.969	578.274	2280.486	0.2535747370734963

由计算结果可见，中文评论里不推荐的样本散度高于非中文评论内部，来说明非中文母语玩家对于游戏中问题的感知更为集中，可能是文化差异所致之类的。而中文评论的批评点明显更为分散，

- UMAP
- 呈现并对比多种聚类方法的结果

总结