Data Collection Methodology

Classes

W. Zhou et. al found that common classes within app review research are “bug reports”, “feature requests”, and “Other.” These can be used for our classification approach as they are vague enough to allow for a myriad of games unbiased toward any genre, these labels also allow for further classification if needed.

Platforms/Data Scraping

Feedback can be collected from Steam as it provides the highest number of possible titles [3] and the highest user count over potential alternatives, hosting over 72,000 game releases since 2004 [3], with a peak concurrent player count of 33 million [4]. Many Steam discussions (not all) separate discussion threads into individual categories, in these cases Bug Reports are typically given their own area, meaning that the data is already labelled as Bug Report/Other. Only games that have these categories can be used to obtain data, as without them data is largely uncategorised/labelled.

Steam is also ideal for HTML scraping as it provides discussion/feedback data in a list format. Scraping is required for Steam as their API does not allow users to pull discussion threads directly. To clarify, Steam does allow HTML scraping within their Robots.txt file.

Metadata such as the Original Poster (OP), date-of-entry and reply count should be included in data collection to allow for additional sorting at a later date. Scraping can be done in Python using BeautifulSoup4 and Request libraries, data can be stored in a CSV file.
Edited to add: One large potential problem with Steam is that the aforementioned discussion list only shows the most recently active topics, meaning that even if thousands of discussion threads exist, without access to Steam’s database they cannot be accessed automatically. Steam still provides a large amount of data that can be scraped, but its not as much as I originally believed.

Language

Steam discussions are not region-based, as such there is often a variety of languages in any game’s discussion section. For additional data verification/validation, a researcher may need to read the data to ensure it is in the correct category, as the researcher (Callum Hemingway) is monolingual, discussion threads written in other languages should be filtered out for this purpose.

Note to supervisors: would using translation software be okay or would it be too inconsistent/not provide an accurate image of what the data contains? Im going to look into this.

Sources

[1] W. J. Martin, F. Sarro, Y. Jia, Y. Zhang and M. Harman, "A survey of app store analysis for software engineering", *IEEE Trans. Software Eng.*, vol. 43, no. 9, pp. 817-847, 2017.
[2] W. Zhou, Y. Wang, Y. Qu and L. Li, "Automating App Review Classification based on Extended Semantic," *2022 9th International Conference on Dependable Systems and Their Applications (DSA)*, Wulumuqi, China, 2022, pp. 106-115
[3] Statista, Number of games released on Steam worldwide from 2004 to 2023 YTD. Online. Available at: https://www.statista.com/statistics/552623/number-games-released-steam/
[4] Statista, Number of peak concurrent Steam users worldwide from 2015 to 2023. Online. Available at: https://www.statista.com/statistics/1330211/steam-peak-concurrent-players/