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 a designated area, meaning that the data is already labelled as Bug Report/Other (though the specific labels change between each product.) Only games with these categories can be used to obtain data, as feedback for games without these categories is largely uncategorised or 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.

Game/Product Selection

An undesirable aspect of using Steam for data collection is that inactive discussion threads are not present in any direct list/index, a choice made to prevent “load issues” (see **Appendix A**.) Because of this, a larger selection of games must be used to create a dataset of the required size.

To maintain variety in the content being scraped from Steam, multiple genres of games should be used to create the data set, for consistency, these genres have been selected beforehand: “Horror”, “Survival”, “Shooter”, “Puzzle”, and “Sandbox”, these were selected as these genres often having their unique and different problems compared to each other. (source needed, Cal)

To ensure enough data is collected, no less than three games will be collected for each genre, with no less than 50 threads being present for each game.

Language

Steam discussions are not region-based, so many languages are often present 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, to facilitate this, NLP models such as the Helsinki translation models can be employed to translate the text content of a discussion into English.

For simple, fast detection of foreign languages, the Python LangDetect library can be used to produce a language code based on text, this language code can then be placed into a string containing the rest of the Helsinki model name and tokenizer name, the string can then be used to automatically select the correct model to translate any given text (though in some cases the language code generated by LangDetect is different from the language code used to identify Helsinki models, in cases such as these the language code must be converted into the correct form before being added to the model and tokenizer name.)

To gain as much data as possible, the translated text should be stored in addition to the native language text rather than replacing it.

Encoding

Encoding of text data should be done in UTF-8 to allow for non-latin script characters such as Cyrillic and ideographic languages.

Manual data verification

General Discussion threads on Steam often include Bug Reports as a consequence of user error, as such, this may lead to Bug Reports being falsely labelled as General data if taken at the face value of where they were found. Due to this manual data verification will likely be necessitated.

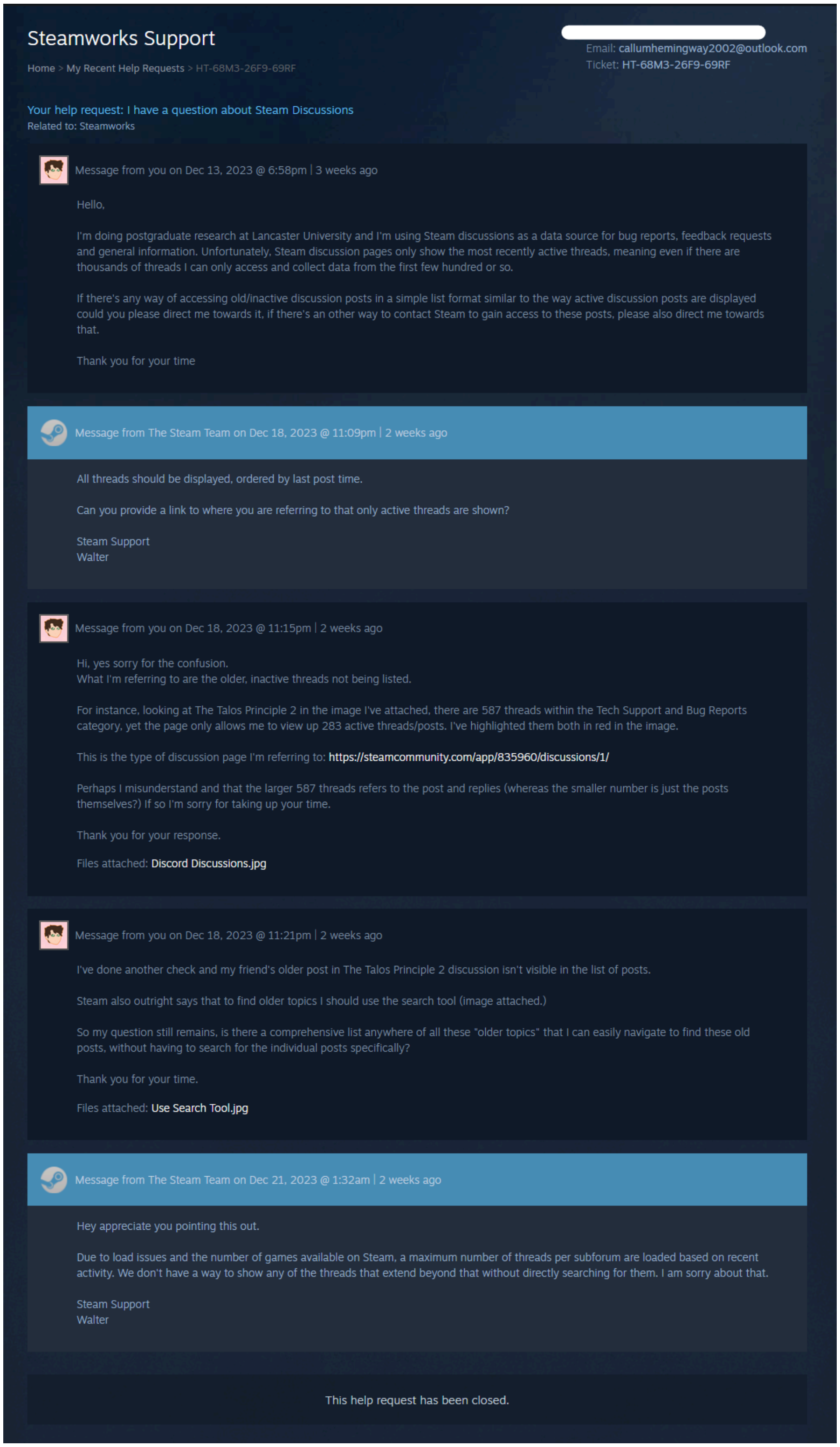
Sources

[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/

Appendix A

Communication with Steam Support agent regarding quick access to inactive discussion threads.



Steamworks Support

Home > My Recent Help Requests > HT-68M3-26F9-69RF

[redacted]

Email: callumhemingway2002@outlook.com
Ticket: HT-68M3-26F9-69RF

Your help request: I have a question about Steam Discussions
Related to: Steamworks

Message from you on Dec 13, 2023 @ 6:58pm | 3 weeks ago

Hello,

I'm doing postgraduate research at Lancaster University and I'm using Steam discussions as a data source for bug reports, feedback requests and general information. Unfortunately, Steam discussion pages only show the most recently active threads, meaning even if there are thousands of threads I can only access and collect data from the first few hundred or so.

If there's any way of accessing old/inactive discussion posts in a simple list format similar to the way active discussion posts are displayed could you please direct me towards it, if there's an other way to contact Steam to gain access to these posts, please also direct me towards that.

Thank you for your time

Message from The Steam Team on Dec 18, 2023 @ 11:09pm | 2 weeks ago

All threads should be displayed, ordered by last post time.

Can you provide a link to where you are referring to that only active threads are shown?

Steam Support
Walter

Message from you on Dec 18, 2023 @ 11:15pm | 2 weeks ago

Hi, yes sorry for the confusion.
What I'm referring to are the older, inactive threads not being listed.

For instance, looking at The Talos Principle 2 in the image I've attached, there are 587 threads within the Tech Support and Bug Reports category, yet the page only allows me to view up 283 active threads/posts. I've highlighted them both in red in the image.

This is the type of discussion page I'm referring to: https://steamcommunity.com/app/835960/discussions/1/

Perhaps I misunderstand and that the larger 587 threads refers to the post and replies (whereas the smaller number is just the posts themselves?) If so I'm sorry for taking up your time.

Thank you for your response.

Files attached: Discord Discussions.jpg

Message from you on Dec 18, 2023 @ 11:21pm | 2 weeks ago

I've done another check and my friend's older post in The Talos Principle 2 discussion isn't visible in the list of posts.

Steam also outright says that to find older topics I should use the search tool (image attached.)

So my question still remains, is there a comprehensive list anywhere of all these "older topics" that I can easily navigate to find these old posts, without having to search for the individual posts specifically?

Thank you for your time.

Files attached: Use Search Tool.jpg

Message from The Steam Team on Dec 21, 2023 @ 1:32am | 2 weeks ago

Hey appreciate you pointing this out.

Due to load issues and the number of games available on Steam, a maximum number of threads per subforum are loaded based on recent activity. We don't have a way to show any of the threads that extend beyond that without directly searching for them. I am sorry about that.

Steam Support
Walter

This help request has been closed.