# DATA301 Project Proposal

Phoebe Zhang 62244228

## Summary

The chosen dataset is Version 2 of the Steam Video Game and Bundle Data. The research question is: which user defined tags are most commonly associated with high user engagement, and can these tags be used to predict the popularity of emerging games? An A-Priori algorithm and a Jacquard similarity will be used to answer this research question. The intended result is to have a list of user-defined tags and tag combinations that are most associated with high engagement games and create a tag-based similarity model that could predict which emerging games may potentially become popular by comparing the user-defined tags to those of previously popular games. The results can help game developers identify game features that attract users and help gamers navigate Steam's vast game catalogue and discover good games before they become popular.

## Motivation

This research question is relevant to me because I enjoy playing games and often use Steam to find new games. I find that when looking for a new game to try out, tags can help to guide my decision, especially when I'm looking for something to match a specific genre, mood or playstyle. However, with the vast collection of available games on Steam, it can be overwhelming to sort through and even harder to identify emerging games that might be truly entertaining and engaging to play before it gains lots of attention.

By analysing which user defined tags are most commonly associated with high user engagement content and identifying patterns which predict the popularity of emerging games, this research forms a direct connection between my own experience as a gamer and Steam user and the skills I've developed through studying algorithms. Sharing the results of this research with others in the gaming industry where everyone is always looking for the something new, it could help them to discover games that they'll enjoy based on patterns identified in community feedback, not purely through marketing.

## Background

The selected datasets contain metadata on available games on Steam and review data of the corresponding game (if there are any). The 'tags' feature in the datasets represents a list of user-defined labels applied to the game chosen by players to reflect their experience, the gameplay and the content of the game. The 'genres' feature in the datasets are broad categories applied to the given game, defined by the developers and are not analysed as a part of the user-defined tags. The 'release_date' feature in the datasets indicates when the game was released allowing the ability to distinguish older, well-known games from newly released ones. The 'hours' feature describes how long the given user has spent playing the specific game, which reflects user engagement.

The A-Priori algorithm is used to find frequent combinations of items (itemsets) and association rules in large datasets. Key features of this algorithm include the support, confidence and lift. The support of an item measures how frequently it appears in the dataset with respect to the total number of transactions. The confidence evaluates the likelihood of two items being together, giving the association between the two items. The lift calculates how much more likely two items appear together compared to if they were independent (compares confidence to support). The Jaccard similarity of sets is the ratio that indicates the similarity between two sets by comparing their intersection size to the size of the union of the two sets.

## Research Question/Hypothesis

The research question is relevant to version 2 of the Steam Video Game and Bundle dataset because user-defined tags, game release dates, game metadata and user reviews are present in the datasets. These features of the datasets make it possible to analyse which user-defined tags correlate to high engagement games and to identify newer games that share similar tags.

The A-Priori algorithm will be implemented to identify frequent user-defined tags and tag combinations in high engagement games (those with high numbers of reviews and high playtimes) and the Jacquard similarity will be computed to measure the similarity of sets of tags between high engagement games and newer emerging games.

## Design and Methods

The dataflow process will begin with loading, cleaning, pre-processing and joining the Steam review and item metadata datasets using Dask. A filter will be applied to the data to include only high engagement games (those with high number of reviews and/or playtimes). From the filtered dataset, the user-defined tags are extracted, and the A-Priori algorithm is implemented to identify frequent itemsets. These frequent itemsets are analysed using the support, confidence and lift features of the A-Priori algorithm to detect associations between tags. The Jaccard similarity is computed to compare the tags of lower engagement or newer games to those of high engagement games. This identifies potential emerging games based on tag similarities to successful, popular games. The results will reveal which influential user-defined tags associate with high-engagement games and potentially sought-after newer games can be recommended established from user-defined tag similarity.

A potential limitation of this research is new games that have little or no user engagement and thus fewer user-defined tags. This will be resolved by setting a threshold that requires games to have a minimum number of user-defined tags. The project aims to find a pattern between user-defined tags and user engagement rather than creating a predictive model. There may be a correlation between tags and engagement, but specific tags might not be the reason for higher engagement in games.

**AI Declaration**

I acknowledge that ChatGPT was used to generate a basic report structure and aid with further explaining technical concepts (A-Priori algorithm and Jaccard similarity algorithms), though all content is my own work, except where I have acknowledged the work of others.

## References


*Apriori Algorithm*. (05/04/2025). GeeksforGeeks. Retrieved 2 May from https://www.geeksforgeeks.org/apriori-algorithm/

Apurva Pathak, K. G., Julian McAuley. (2017). *Generating and personalizing bundle recommendations on Steam* SIGIR,

Jure Leskovec, A. R., Jeff Ullman. Finding Similar Items. In *Mining of Massive Datasets* (3rd ed., pp. 73-134). Cambridge University Press. http://infolab.stanford.edu/~ullman/mmds/ch3n.pdf

Jure Leskovec, A. R., Jeff Ullman. Frequent Itemsets. In *Mining of Massive Datasets* (pp. 213-251). Cambridge University Press. http://infolab.stanford.edu/~ullman/mmds/ch6.pdf

Karabiber, F. *Jaccard Similarity*. Retrieved 2 May from https://www.learndatasci.com/glossary/jaccard-similarity/

Mengting Wan, J. M. (2018). *Item recommendation on monotonic behavior chains* RecSys,

Wang-Cheng Kang, J. M. (2018). *Self-attentive sequential recommendation* ICDM,