COMP 7150 — Data Science Project Report

Hicks, Eric
elhicks@memphis.edu

Kelly, Craig
cnkelly@memphis.edu

December 2, 2016

Abstract

A raw data source was mining for data, which was cleaned and filtered into a usable dataset. This dataset was searched for interesting details based upon the authors' preconceptions. They were found.

1 Introduction

In this paper, the path and results of the authors' dissection of the Steam platform will be detailed. The reason Steam was chosen for this project, was due to its large acceptance in the PC space. Steam is a company that sells PC, Mac, and Linux games for download over the Internet. With weekly sales, daily releases, a managed library, and no computer limit on games, it is the largest digital distributor of games and the most widely used. Additionally, Steam stores metrics on over a 100 million users on nearly 10,000 games, making it a perfect example of a feature rich dataset. However, as Steam has no saved repository of data publicly available, one was created for this project, hereby called the Steam dataset. [1]

2 Steam Dataset

Although, several external sites keep track of various parts of Steam, none of them gathers the data already available by web crawling Steam. Of the few sites that record data not saved by Steam such as Rhekua [2], none of them release snapshots of their gathered data. This lead the authors to manually mine data from the main Steam site.

TODO: appendix with field listing? or just a link to the repo?

2.1 Acquisition

TODO: multiple full pulls from steam and merge with steamspy

2.2 Cleaning and Formatting

TODO: reduced to 78 fields TODO: explain field removal

2.3 Caveats

Some aspects of steam were not able to be captured in the data pulled from Steam. Some are not recorded and others can only be accessed by an administrator account. The number of owners for each game could not be acquired, but has been estimated using an outside site [3]. Recommendations were collected as a total of reviews and as two separate values for positive and negative reviews. The sentiment of the reviews that Steam displays was also not able to be captured. Steam doesn't store historical price data, so sale history and price trends can not be analyzed from the data. These limitations did not stop the authors from gaining interesting insight from the data.

3 Predictions

As both authors were familiar with Steam before the start of the project, they made some predictions of what would be found in the data:

1. SteamOS games available on Steam would be a perfect match to Linux games available.
2. The most recommended and the highest rated genre is action.
3. That Metacritic scores are an inverse bell curve when sorted by recommendation, i.e. lower and higher scoring games would have more recommendations that games with a middle score.

4. That free games get more reviews per game both are rated lower than paid games.

4 Testing

Both authors ran different models on the data independent of each other to cover the most ground. Every feature pair was plotted as a scatter plot to gain a basic understanding of feature relations alongside reading through the raw data. From here other options were tried.

5 Results

The following are the results of the authors' predictions:

1. It was verified that all SteamOS games are also the entirety of the Linux library on Steam.
2. The most recommended genre was free to play and not action. The least recommended genre was non-game software.
3. The highest scoring genre was sports instead of action. The lowest scoring genre was free to play.
4. Free games do get more recommendations than paid games, they also are rated lower than paid games.

The above leads to the conclusion that free to play games likely have mostly negative reviews. These findings also lead to the idea that free games receive mostly negative reviews. Other results pulled from the data not related to the authors' predictions include:

1. Pricing compared to Metacritic scores is mostly uniform. The gaps for certain prices are almost entirely due to Steam's pricing structure.
2. Pricing compared to user recommendations is also nearly uniform. There is a small increase in recommendations for cheaper games, but it is not significant.

6 Future Work

Given a larger timespan to collect data and more sophisticated web crawlers, more data come be pulled from Steam. This would allow for most, if not all of the caveats mentioned before to be removed. With historical data, price trends for games of different genres could be predicted along with the time frames for price drops. A divided recommendation statistic could be used to determine if certain genres get more positive reviews than others, as well as show a correlation (or lack thereof) between Metacritic scores and positive recommendations.

7 Conclusion

After finding a suitable platform to mine data from, the authors acquired raw data from Steam. This data was cleaned, formatted, and processed to find several interesting results. While all of the authors' predictions were not proven in testing, the ones that were wrong proved to be the most surprising.

References

[1] Valve Corporation. *Steam*. URL: `http://store.steampowered.com/`.

[2] Dillon DeLoss. *Rhekua*. URL: `http://steamsales.rhekua.com/`.

[3] Sergey Galyonkin. *Steam Spy*. URL: `http://steamspy.com/`.