Data Engineering and Analytics Project Concept

Kirchebner Paul, Kommer Phillip, Pickelmann Florian

9. November 2022

1 Introduction / Research question

Like most people nowadays we have a high affinity of everything pop culture related. So of course we wanted a topic which is in some way or another related to that. We decided to focus on gaming because there is a lot of information and statistics out there which we can use. So we thought about what question we personally would like to answer:
Can one predict the popularity of a computer video game based on its tags. To answer this question we will look at the tags and the reviews of games on Steam.

2 Background and challenge

The world of video games is always changing. A few years ago MOBA games like "Dota 2" or Ego-Shooter games like "Counter Strike" dominated the market but nowadays so called "Battle Royals" like "Player Unknown Battlegrounds" are leading the popularity leader boards. But just because they are popular doesn't mean that they are well liked by a lot of people. So we would like to analyse all genres and sub-genres that are out there and see if you can make the "perfect" game which should in theory be popular and also well liked by most people. This shouldn't be an easy task because there are a lot of genres out there and an always changing player base.

3 Correlation to other works

One correlated work would be a 2020 paper from Xiaozhou Li and Boyang Zhang called *A Preliminary Network Analysis on Steam Game Tags: Another Way of Understanding Game Genres*. In this conference paper the two students from the Tampere University in Finnland analysed co-occurrence of tags in games on Steam. They did this in order to find connections and the distribution of the game tags.

Another related work is an article from Nick Yee, written for Quantic Foundry, which is a market research company focused on gamer motivation. In his 2018 article *Visualizing How Steam Tags Are Related* he analysed the tags of 2,129 games and visualized the results to see how often the tags appear and if there are relations between them.[1]

One 2016 paper printed in the 54 Vol. of the *Cataloging & Classification Quarterly* journal also covers the topic of tags in Steam games. The paper *Full Steam Ahead: A Conceptual Analysis of*

[1] https://quanticfoundry.com/2018/01/24/visualizing-steam-tags-related/

UserSupplied Tags on Steam of the four authors Travis W. Windleharth, Jacob Jett, Marc Schmalz and Jin Ha Lee looks at the concept of user-generated tags and analyses how this system in Steam plays out. They categorized the tags and tried to identify useful metadata and terms for future work.

4 Required data

We need a lot of different games with different tags and for each one of them their popularity and a metric how well liked it is. Fortunately, the biggest PC online games distributor platform "Steam" [2] has all the data we need. From this website you can see how many people own this game, how popular it currently is (how many player are playing it right now) and how the users are liking this game, based on their reviews. The upper mentioned tags can also be retrieved for Steam. The tags are generated in collaboration with the users, but due to the enormous size of Steam they should be quite accurate.

5 How do we retrieve this data?

Steam has its own API [3] which already provides us with a lot of the data we need. Unfortunately the amount of requests per minute is restricted but for our purposes it should work out, because we will probably not include every title ever released on Steam. Because some of the data we need isn't easily obtainable from just the Steam API, we also use an API from a website called "SteamSpy". [4]. Steamspy is also getting their data from the Steam API but they have been collecting Steam data over a longer time period and keep updating their data. The requests on their API are also limited but with enough time and combined with the Steam API we will be able to get all the data that we need. Due to SteamSpy's data coming directly from a trustworthy source, Steam, we can ensure that the quality of the data is sufficient for our purpose.

6 Methods and resources for the future

Due to us looking at many factors (tags, reviews, player count, ...) we will start with a Multivariate DEA. We are not planning to focus our whole project on visualization, like Nick Yee did it in his article, but we will probably visualize at least some data for better comprehension. On the one hand, we will analyse the tags quantitatively (which are to most/least popular etc.), but on the other hand we are planning a similar approach to Xiaozhou LiTampere and Boyang Zhang in their paper and construct a network out of the tags. We will then apply some method of centrality measures to the tags to better analyse the network of tags and the relations between tags. The gathered information will then be combined and evaluated with the review and player count data to find connections.

[2] https://store.steampowered.com/?l=english

[3] https://steamcommunity.com/dev?l=english

[4] https://steamspy.com/api.php