Jonah Melhado

07/31/2020

Final Capstone Report

Brain Station 2020

Steam Game Recommender

Problem Statement:

Games only last so long and so gamers are constantly looking for new content to fill their valuable playtime. Most gamers are also looking for the best possible game on which to spend their limited funds, but oftentimes mainstream recommenders are not adequate, and don't allow the user to target their recommendations appropriately. This project aims to fill that gap by allowing users to apply various filters to the recommendation process in order to tailor the process to their needs.

Background:

Data science techniques are perfect for this problem as they will allow the creation of different kinds of filters which will intake various content features of the applications and preferences of the users who play them to curate a set of games that the user is more likely to enjoy. In the past this problem has been addressed using several different types of filtering, including user-independent systems, content based filtering and collaborative based filtering. Each method has its own intricacies as well as various different styles of application. Although these methods have been used in the past, there is limited customization available on the part of the user which is where new data methods could improve the quality of the recommendations by allowing for greater user input.

The Data:

Part of the data used in this project was sourced from Kaggle. The data set was created using data gathered from the Steam Store and SteamSpy APIs. SteamSpy is a website created by an independent steam user that contains various metrics on games in the Steam Store. The data set contains the following columns describing over 25,000 applications:

Column Name	**Description**
appid	Unique Steam Store identifier for each game
name	The title of the game
release_date	The release date in the format YYYY-MM-DD
english	Boolean column indicating whether or not English is a supported language
developer	Name(s) of developer(s)
publisher	Name(s) of publisher(s)
platforms	Operating systems that are compatible
required_age	Minimum required age in the UK

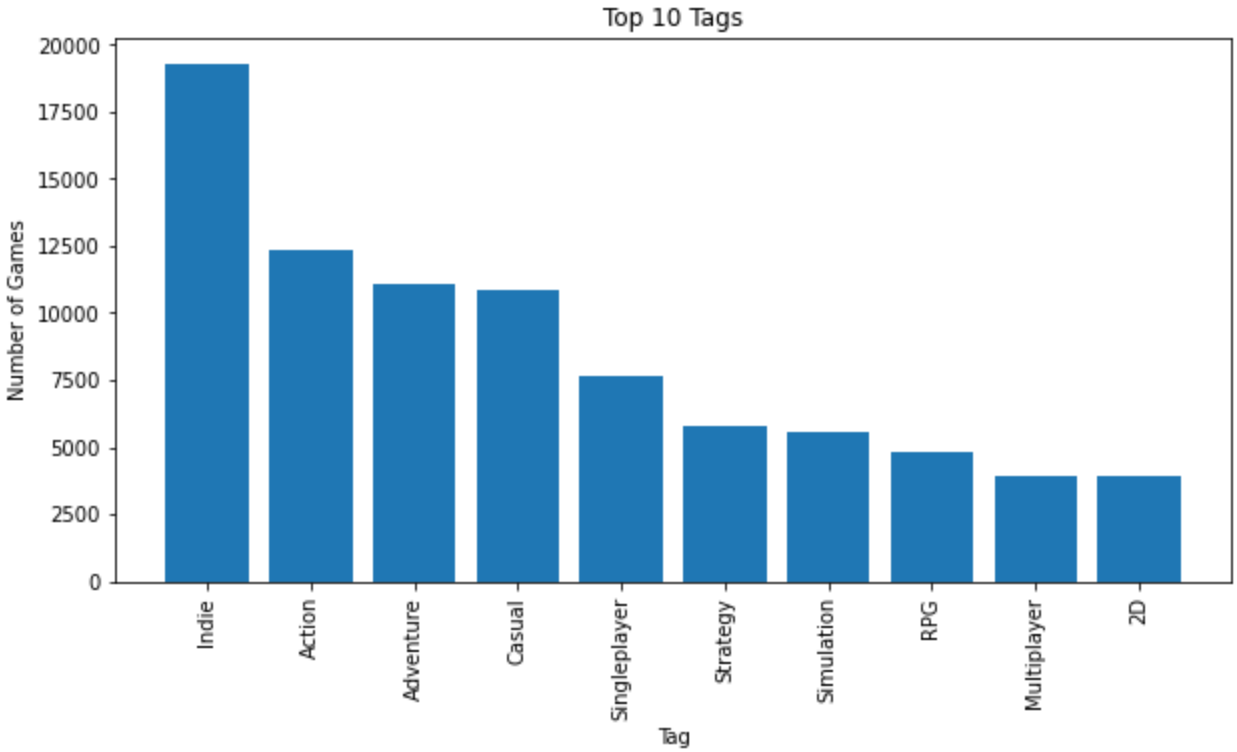
categories	List of game categories
genres	List of game genres
steamspy_tags	List of top steamspy game tags
achievements	Number of in-game achievements
positive_ratings	Number of positive ratings
negative_ratings	Number of negative ratings
average_playtime	Average user playtime
median_playtime	Median user playtime
owners	Estimated number of owners
price	Full price in GBP

The other part of the data which included individual user reviews and hours played for each of the applications was independently web scraped. The web scraping was an extremely time-consuming task even without consideration for time spent debugging and refining the process, taking approximately seventy-two hours to complete scraping millions of user reviews. This may seem like an excessive quantity of time; however, this time was only achieved due to optimization of the scraping function so as to eliminate the need for excessive cleaning after the fact.

The webpages containing the reviews were infinite-scrolling and therefore did not contain individual pages of a set number of reviews, which would have allowed the use of the requests package to simply loop through each page and collect all of the review data. Considering the behavior of the pages, it was necessary to use a package called Selenium which allowed a driver to independently control a browser instance to continuously scroll until the entirety of the body of reviews was loaded onto the page before extracting the desired data and saving it as a large batch of CSV files. Unfortunately, upon completion of the data collection there was not enough time available to process the acquired data into a usable format due to its size. Even so, upon examination it was easy to see that there was not a great amount of overlap in the usernames collected by the scraper for each game, indicating that not enough data was acquired about each user for representative vectors to take form.

Final Summary:

Due to time constraints, the recommender created was a simple Streamlit browser application which recommends games based on their user-defined content by intaking a single game and, if desired, a category which allows the user to tune the recommendation results slightly. The game descriptions were scraped from the Steam Store from the user-defined tags. These descriptions are not completely representative of the contents of each game; however they are the most accurate descriptions available. Below is a graph which shows the top ten tags by frequency of appearance.



As was initially expected the most frequently encountered tag was ‘Indie’ which indicates that the game was developed by a small independent studio or developer as opposed to a well-known content creator, with ‘Action’ and ‘Adventure’ coming in closely behind. Although this product is unfinished, after allowing ten individuals to review the recommended games, it was deemed adequate for those looking for a general direction for their next purchase, however it requires much refinement.

Going Forward:

The end goal of the recommender remains as it was from the beginning. The aim is to create a web application that allows a user to log in with their Steam account and input parameters for desired recommendations including genre, release date, range of hours necessary for completion, among other desirable metrics. In order to allow for the manipulation of a large user review dataset, AWS will be employed in order to avoid computing limitations and a more thorough data gathering process must be put in place. A possible path which will be further explored is examining the user-base itself rather than going application by application. This would allow for the acquisition of multiple reviews per user as long as their profile is visible to the public. To further strengthen future iterations NLP will be employed to more thoroughly grasp the aspects of each game which users enjoy and thereby increase the ability to cluster users based on their preferences. There are still many directions that can, and will, be explored.