Steam Scraper Project Proposal

Rob Bray, Matthew Dixson, Tom Hills, Tan Nguyen

***Abstract*— Video games are becoming increasingly popular. We will explore how much a video game's genre affects different aspects of the game, such as rating, price, discounts, extra downloadable content, and computer requirements.**

I. Objectives

Our objective is to scrape data from Steam's game library to gain information about genre, price, sales, review ratings, and other metrics from all listings. Genre would be analyzed for game quantity as well as how it affects the word choice in the game's description. Price and sales figures would be collected to give average price per genre as well as total sales. The other metrics would have similar analysis performed to create a complete statistical picture of Steam's general library.

Steam games tend to go on sale quite often, so we are also interested in seeing how many games in a genre are on sale and what the average sale percentage for a genre is. Extra downloadable content (DLC) for games is also sold on Steam. We also want to determine if some genres have more DLC than others. Steam also allows users to rate video games they have purchased. It would be interesting to see which genre has the highest and lowest average rating.

II. Motivation

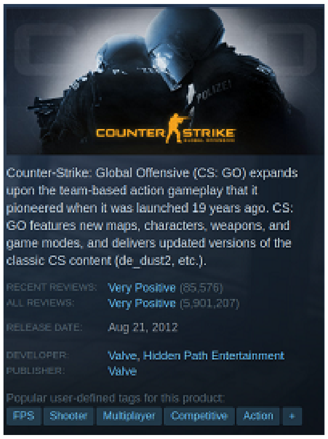
Steam is one of the largest retailers for video games on PCs. Steam also sells games for a discounted price quite often. Knowing when games go on sale, how much their price is discounted when they go on sale, and knowing a game's user ratings are important to video game enthusiasts, because these things influence if they will purchase a game or not. It would also be beneficial to buyers to know this data and more by video game genre. With this information, a buyer can look in the genre with high average rating and large discounts for highly rated video games that are on sale.

III. Data Obtained

On each game page, Steam includes a lot of information about the game, like its description, genre, average ratings, DLC, and system requirements. We are interested in these pieces of information for several types of analysis. Firstly, we want to analyze how different genres use words to advertise their games. This includes the frequency of certain words and the average length of words.

We are also interested in the DLC the game has. Each Steam game page has a section for the game's DLC if it has any (Figure 2). There has been a long-running debate about whether or not DLC is good for the gaming industry. We would like to compare average ratings to the number of DLC that a game has and determine if there is a correlation between the two.

Lastly, we would like to compare the game's genres to the game's system requirements. Each Steam game page has a section that includes its minimum and recommended system requirements which are the recommendations for making the game playable (Figure 3). We think that some genres may yield themselves to better-looking graphics, so we would like to see if this is true. This part is tricky, since system requirements are not simple numbers. Each component has a unique name and number. However, we could use the name of the component to search for benchmark numbers to use that for our analysis. The benchmark number is an arbitrary number assigned to components based on how they perform compared to other components.



Counter-Strike: Global Offensive (CS: GO) expands upon the team-based action gameplay that it pioneered when it was launched 19 years ago. CS: GO features new maps, characters, weapons, and game modes, and delivers updated versions of the classic CS content (de_dust2, etc.).

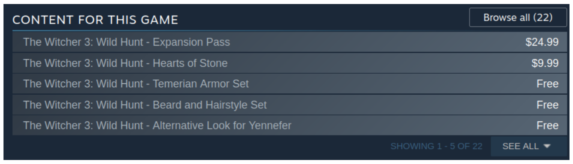
RECENT REVIEWS: Very Positive (85,576)
ALL REVIEWS: Very Positive (5,901,207)

RELEASE DATE: Aug 21, 2012

DEVELOPER: Valve, Hidden Path Entertainment
PUBLISHER: Valve

Popular user-defined tags for this product:

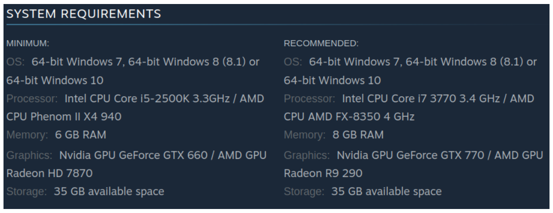
Fig. 1.



CONTENT FOR THIS GAME

The Witcher 3: Wild Hunt - Expansion Pass	$24.99
The Witcher 3: Wild Hunt - Hearts of Stone	$9.99
The Witcher 3: Wild Hunt - Temerian Armor Set	Free
The Witcher 3: Wild Hunt - Beard and Hairstyle Set	Free
The Witcher 3: Wild Hunt - Alternative Look for Yennefer	Free

Fig. 2.



SYSTEM REQUIREMENTS

MINIMUM:
OS: 64-bit Windows 7, 64-bit Windows 8 (8.1) or 64-bit Windows 10
Processor: Intel CPU Core i5-2500K 3.3GHz / AMD CPU Phenom II X4 940
Memory: 6 GB RAM
Graphics: Nvidia GPU GeForce GTX 660 / AMD GPU Radeon HD 7870
Storage: 35 GB available space

RECOMMENDED:
OS: 64-bit Windows 7, 64-bit Windows 8 (8.1) or 64-bit Windows 10
Processor: Intel CPU Core i7 3770 3.4 GHz / AMD CPU AMD FX-8350 4 GHz
Memory: 8 GB RAM
Graphics: Nvidia GPU GeForce GTX 770 / AMD GPU Radeon R9 290
Storage: 35 GB available space

Fig. 3.

IV. Responsibilities of Team Members

All Members - scraping each component of data. Since scraping is a big component of this project, we would all like to learn how to do it properly.

Tom Hills - analyzing price, and word frequency and length of words in description, comparing to genre.

Matthew Dixson - analyzing number of DLC packages and comparing this to the average ratings

Rob Bray - analyzing minimum and recommended system requirements and comparing these to the genre

Tan Nguyen - Comparing minimum and recommended system requirements to each other, analyzing the disparity between the two in different genres

V. Milestones

1) October 11: Be able to get HTML from the Steam store programmatically. Begin looking at how to parse the HTML for data

2) October 18: Create an organized database based off of results from the previous week

3) October 25: Have a web scraper that can take a url to a Steam page and get the information need to make an entry in the database

4) November 1: Begin analyzing the data that has already been obtained. Each member will focus on the analysis assigned to them above. Start looking at ways to gather large amounts of data from Steam at once. For example, get data from 50 games of a certain genre.

5) November 8: Use the scripts we wrote for last week on the larger dataset. Adjust the scripts as need be

6) Final report due date: Start the final report 1-2 weeks before the deadline.

VI. Expected Outcomes

A. Genre-Description Analysis

We expect to find significant correlations for certain types of words for each genre analyzed. For example, we expect action and adventure games to use verbs and explosive language more often than point-and-click story games.

B. DLC-Ratings Analysis

We expect to find that games with more DLC will have higher ratings than games without DLC. DLC adds more content to a game which allows fans of the game to play it more. DLC can also fix problems that exist with the original game by introducing new features that the original lacked.

C. System Requirements - Genre Analysis

We expect to find that more intense genres (action, adventure, horror) have higher system requirements than more laid back games (cozy, storybook), since these games are more about the intensity of the experience.

D. Minimum-Recommended Requirements Analysis

We expect to find that more action-oriented titles will have a larger disparity between their minimum and recommended requirements, since these usually make up the majority of the game market.

E. Price Analysis

We expect that older games and games from smaller developers will have a lower price. New games and games from bigger development studios will most likely have larger price tags. The amount of paid DLC might also influence the price of the original game. Paid DLC might lower the price of the original game to incentivize buying both the game and the DLC.