Tristan C

INFO 5810.003

Summer 2021

# The Paradigm of Video Game Reviews:

# An Analysis of Reviews for the Game “The Last of Us”

## About Data and Source

The dataset that I will be utilizing for this assignment is related to one of the most growing and dynamic industries around us, that being the video game industry. This dataset in question was set out to do one specific thing, which is to scrape and curate all of the review data for a very specific game and sort it out between critics that have reviewed the game and the users themselves that have reviewed the game.

The game in question that we are using from the dataset is the PlayStation exclusive game “The Last of Us”. Released in 2013, The Last of Us has been considered one of the better games of the later PlayStation era of games and has become such a popular icon enough to where it is even getting its own HBO show that is currently in production. The goal of this dataset that the creator had in mind was to curate all the reviews for this game from the popular website Metacritic. Using a webscraping tool named Scrapy and Beautifulsoup, the creator was able to download a total of over 5000+ reviews from users and critics alike for the game. For this specific dataset, I will be focusing on the userbase reviews, as they contained the bulk of the reviews (4000+) and are the most varied to give us some more interesting results.

## Goal of Analysis

Within the dataset that we are given, there are a few goals that I have in place to try to solve whenever utilizing and analyzing the data in question. Some of the goals are as follows:

* Attempt to find and curate and patterns within the dataset that could describe the consensus that comes with the reviews
* Find out if there is a difference between people who gave normal reviews versus extended reviews.
* Determine commonalities amongst reviewers to see what is the most loved and hated aspects of the games based on common words in reviews.
* Determine if the ratings differ between the two consoles that the game was released on (PS3 and the remastered PS4 version)

## Data Mining Techniques

The following data mining techniques will be used to help determine our analysis goals:

* Correlation Analysis – A statistical measure that will calculate the data results based on the similarity of the set value/attributes in a data set. The results can either skew towards a positive number, implying that the attribute’s values go up or down as does the correlated results or negative which implies the results move in opposite directions.
* Textual Association – This method of data mining focuses on taking unstructured text-based data and using association to display similar words via confidence levels as an example. This can be used by many to help discover patterns of words or find commonalities between each section.
* K-Means Clustering - This method helps to enables the user to define and curate groups between data sets by comparing the means (or average) of their individual values. After curating the average values, the items are then sorted into “k” type clusters and sorts them in a way that each cluster is closest to their nearest mean value.

## Results of Analysis

* Correlation Analysis – This is one of the few that I was not able to really get a definitive answer on. This one was going to help me with the goals of seeing if the reviews differed from platform to platform. While looking at it, it seems that the type of review (normal or extended) or the platform (PS3 or PS4) really did not make too much of a difference, with the review score and platform having such a minor -0.004 difference. Despite that, I can still say in a way it is interesting to see that it really did not matter which platform or console the game was played on. The game itself still received universally identical ratings across the board, which in a way could be used to prove that each version of the game could be played to completion, and you are guaranteed to have a similar experience no matter what the limitations may be.
* Textual Association – When searching results, there were a lot of associations that could be made, even with a high level of confidence which shows that a lot of the words in question do strongly correlate with one another. A lot of the words that correlate involve the word story and some sort of positive term such as masterpiece, great, good, etc. which would easily give us the indication that the story is one of the strongest things that people loved about this game. Gameplay and characters followed the same suit. Naturally, there is also the context that the word “game” is heavily correlated as it is constantly mentioned and I do think causes some bloating with the rules, so with further adjustments, we could get a more concise group of results but the results/data itself is still the same.
* K-Means Clustering – The first thing that immediately caught my eye is that the way that the clusters decided to divvy themselves up is related to similar aspects to a certain degree. Each cluster is basically related to a certain type of feeling of how the games made the user feel (such as this part was good, great, masterpiece, amazing) etc. and the cluster defined themselves on how close each thing was. Cluster 0 is about aspects people considered great, cluster 1 is about aspects people considered amazing, cluster 2 is about what people think is a “Masterpiece” and is a much more summative statement on the game itself, and finally cluster 3 is about what people considered just good. It is quite interesting to see that is how the clusters decided to split things up, but it retrieves some interesting results from all of it.

## Diagrams

### Correlation Analysis

A screenshot of a computer

Description automatically generated with medium confidence

### Textual Association

A screenshot of a computer

Description automatically generated with medium confidence

### K-Means Clustering

Diagram

Description automatically generated