# Steam Reviews Frequent Words

Gibson, Samuel

## Introduction

Steam is a video game digital distribution service and storefront created by the company Valve. Video game publishers can add their game to this online store for users to purchase, download and play. As of 2022, the platform contains 62.6 million users and includes 50,361 games available for purchase and download. Clearly such a popular platform will be very lucrative for video game publishers and developers. The focus of this project will be to analyze the bountiful product review feature of Steam for words that can potentially aid in product marketing and development.

## Statement of Task

The task of this project will be to discover which words are frequent in both positive and negative reviews on Steam. The question this result will aim to answer is, "what key words are determinants of a successful video game product?". For instance, the average popular video game on Steam could be attributed to words such as "fun" whereas a video game with poor reception could be attributed to words such as "bad". Though, simply discovering words such as "fun" or "bad" is not significant. To retrieve significant and interesting results, these simple obvious words need to be disregarded. The challenges of this project are to retrieving non-obvious results, data preprocessing and cleaning, and data sampling decisions.

## Dataset

The dataset used is a collection of 6.4 million English reviews from the Steam reviews portion of the Steam store. The dataset provides the product name, id, review text, the review score, and the review votes. Name, id, and review text are self-explanatory. The review score is an integer representing whether the review was positive or negative with 1 being positive and -1 being negative. The review votes are a Boolean value of 1 or 0 indicating whether another user recommended the review.

## Methodology

Upon inspection of the initial dataset, it is very dirty as well as being massive. To clean and preprocess the data into a useful form several methods are employed.

First, to address the size of the dataset, only a sample of it is used. This is done to save on performance and time. Results were obtained using a sample of 1%, 10%, and 35% of the data. Considering the original size, these samples should prove sufficient since we are just interested in frequent words.

Cleaning the data requires more steps to achieve useful information. Steam reviews allow the use of special characters, and often users will take advantage of this to create text figures in reviews. These reviews are not helpful and therefore all special characters need to be removed. As with any text, punctuation, casing, and forms of lemmas are present. Additionally, duplicate reviews are also present in the dataset.

All this noise can produce bad results which will need to be avoided with the use of several processes. To begin, duplicates are dropped from the dataset followed by separation of positive and negative reviews. Then each word in each dataset's reviews text is retrieved via tokenization. The library NLTK provides a word tokenizer which will be utilized. Once the words are tokenized, lower casing is enforced onto each word as well as removal of punctuation. Punctuation is removed by creating a mapping table which will remove all punctuation characters provided in Python’s string.punctuation list. Special characters are next removed by simply checking each word in a review to check if it is not a special character. This is done by utilizing Python’s isalpha() method. To retrieve significant frequent words, stop words are next removed. Stop words are insignificant words that frequently appear in text and language. For example, these words could be “the” or “I”. NLTK provides a collection of these words which is used to filter out stop words from reviews. Since these reviews are for video games additional stop words are added to the filter list, such as “game” and “play”.

Normalization is the next process utilized in cleaning the data. Since words can have several forms, these can be reduced into their base form to increase the accuracy of results. Two strategies were considered for normalization, stemming and lemmatization. Stemming utilizes the stem of a word whereas lemmatization involves context analysis. The process for lemmatization is more computationally expensive, especially for large datasets, but it results in greater accuracy. Since the dataset is sampled, lemmatization is possible. Another advantage of lemmatization is that it can identify proper nouns. Since reviews sometimes contains the name of another game, it would be significant if it could be a frequent word. Stemming would have trouble when met with these proper nouns, especially since we have already preprocessed the reviews.

To normalize the reviews, NLTK’s pos_tag and WordNetLemmatizer are utilized. Each word in each review has its type tagged with values such as noun or verb. This enables the lemmatizer to analyze the context for words in reviews and properly reduce them to their base form. The simple algorithm is presented as follows:

```
$for\ word, tag\ in\ review\ sentence$
$if\ tag = form\ of\ noun$
$pos = n$
$else\ if\ tag = form\ of\ verb$
$pos = v$
$else$
$pos = a$
$results += lemmatize(word, pos)$
```

This will determine the word’s type according to its tag, then run NLTK’s lemmatizer on the word to replace it with its base form. The result is the review text lemmatized. The final step for processing the data involves organizing words by frequencies and filtering out non unique words. Filtering out similar words between positive and negative review words is important to avoid meaningless words. If a word is shared between positive and negative reviews, then it has little meaning when it comes to review reception and is most likely a stop word that slipped through the data preprocessing. The top frequent words are then gathered for positive and negative reviews and reflect user sentiment.

## Results

Data was processed and then plotted into a word cloud for viewing for each dataset sample amount. The top 30 frequent words for negative and positive reviews were used prior to removing common words.

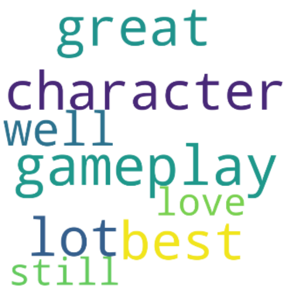

Image 1: *Frequent words wordcloud associated with positive reviews.*

| Sample Size | Positive Words | Negative Words |
|---|---|---|
| 1% | Best, gameplay, love, well, great, character, lot, hour. | Look, try, way, give, money, bad, work |
| 10% | Gameplay, love, lot, great, well, still, character, best | Look, bad, way, work, try, money, give |
| 35% | Great, character, well, gameplay, love, lot, best, still | Look, use, way, try, give, bad, money |

Table 1: *Frequent words by sample size and reception*

Sample size used proved to have little difference in the frequent words, only noticeable changes being the words “hour” and “use”. Some of the words have vague meanings, such as “well” and “way” therefore providing little useful interpretations. The existence of these words could signify some of these should be considered stop words. Despite this, there were a few significant words associate with each category. Gameplay is an obvious result for positive reviews, since users who enjoy playing a game would state this in a positive review. Character was an unexpected word, and therefore the most significant for positive reviews. Video games are often defined by the characters involved, either as storytelling pieces, or by simply having a unique and attractive design. The existence of this word signifies that users value well developed characters in videogames and are likely to leave a positive review on a game that contains these. As for negative reviews, the words “work” and “money” are significant. Money could indicate that users poorly review games that cost more than they believe it to be worth. The existence of “give” as a negative word could also compliment “money” as this could signify that users regret giving their money to the specific video game publisher. This is difficult to interpret since its meaning could either lean towards the cost of a game, or the user’s dissatisfaction of spending money on the game. As for the word work, this could indicate that users dislike video games that make them feel as if they need to do a lot of work to play the game. This could relate to the difficulty, tediousness, and repetitiveness of a game.

The positive word results more clearly answer the project task question with the word “character”, whereas the negative results do not. The negative words are less clear in meaning but can still be defined given the context of videogames.

## Conclusion

The results provided by the methodology proved enough to provide a somewhat substantial result. Though the number of useful words retrieved were not high, they are still useful. Video game developers looking to publish a game onto Steam could emphasize character design and gameplay should they want good reviews. To avoid bad reviews, ensuring their game doesn’t make users feel like they are working and properly pricing their product are potential strategies. These words can also be used for marketing a product as well as knowing which words to avoid.