Professional Videogame Reviews

Gonçalo de Abreu Matias
Faculty of Engineering - University of Porto
up202108703@up.pt

Henrique Filipe Pereira da Silva Caridade
Faculty of Engineering - University of Porto
up202108817@up.pt

Pedro da Cunha Teixeira e Faro Beirão
Faculty of Engineering - University of Porto
up202108718@up.pt

Pedro Miguel Marques da Silva Pinto
Faculty of Engineering - University of Porto
up202108826@up.pt

Figure 1: Gaming Setup

Abstract

In recent years, the video game industry has revolutionized the way we live, and the way we play. In order for people to be able to quickly have a opinion about some game, appeared the professional reviews, that have a big influence and can even impact sales. With this project, we are addressing the need for an efficient system that allows users to quickly search and check reviews about potential game they might be interested in buying for example, or only searching for a second opinion about a game they played. With this article, we are documenting the methods we used to implement the Professional Video Games Reviews search engine.

Keywords

Video Games, Opinions, Gamers, Professional, Reviews, Information Retrieval, Dataset, Data Collection, Data Analysis, Search Engine, Review Aggregation, Professional Criticism, User Experience,Platform Comparison

G65, October, 2024, Porto, Portugal (FEUP)

ACM ISBN 978-1-4503-XXXX-X/18/06

ACM Reference Format:
Gonçalo de Abreu Matias, Henrique Filipe Pereira da Silva Caridade, Pedro da Cunha Teixeira e Faro Beirão, and Pedro Miguel Marques da Silva Pinto. 2024. Professional Videogame Reviews. In *Proceedings of PRI (G65).* ACM, New York, NY, USA, 5 pages.

1 Introduction

We are developing this search engine for the Master's in Informatics and Computing Engineering (PRI - MEIC - FEUP), in groups of four students.

While growing up, gaming has always been part of our lives, and that is why we think we couldn't choose any other topic for this project. We are aware it is a really rich and diverse industry, with vast amounts of data to explore, and we think it would be a nice addition and a very useful tool for people searching for gaming reviews online.

We are dividing this report into multiple sections, to provide a clear idea of what we are implementing. We start by describing why we chose this dataset, then we describe how we went about getting the dataset and the difficulties we encountered along the way. After that, we show the characteristics of the final dataset and finish by proposing some search scenarios we expect to be able to provide for our users.

2 Data Preparation

The first thing we were tasked with was to get a dataset with unstructured data and a pipeline that allowed us to extract more if need be, which we would then use in later stages of the project. This process varies wildly from topic to topic, but in our case, we didn't find any good existing datasets, so we decided to scrape from a website with game reviews.

2.1 Data Selection

After we decided that we wanted to scrape a website for our dataset we had to decide which website it would be. As there are many game review websites we searched through many of them to find one we were happy with but in the end, we landed on IGN [4]. IGN [4] is a well-known website that has vast experience with video game reviews. It is one of the oldest in the business of video game reviews having reviews dating as far back as 1996 so after we found it, it was a no-brainer to have IGN [4] as our source of game reviews.

3 Data Pipeline

The problem with IGN [4] is that it doesn't have an API and doesn't have a place where information about every review is displayed. The only way to get the review's contents is by the review's URL, and to add insult to injury, it doesn't have a place to get all of the reviews' URLs.

The full pipeline as described in more detail ahead is illustrated in Figure 11.

3.1 Selenium Extraction

With that said, the first step in our pipeline is to get as many review URLs as possible. The best way we found to do this is with Selenium [3] because IGN's [4] game review page uses a technique known as Infinite Scrolling which means the user needs to scroll before the website loads more data, Selenium [3] is perfect for that because it allows you to mimic a user and scroll to the bottom of the page.

With Selenium [3] we just let it run for a while but found that either ign[4] or Selenium would fail at around 150th scroll and would crash the Selenium script. So we used the score filter and started by scraping the reviews with a score of 10 and the reviews with scores between 9 and 9.9 until 0. We found that this way, we could get more URLs by running it with different score ranges and 100 scrolls as the limit (most of the average score ranges got to this limit) and then end it by saving the page HTML source code to a file.

We didn't get all of the scores available on the website but we got a good part of them.

3.2 URL Extraction

The second step is to extract the review URLs from the HTML source code for all the score ranges. For that, we used BeautifulSoup [6].

BeautifulSoup [6] allows for easy traversal of the HTML document tree in Python [8]. With it, we were able to extract URLs embedded in the HTML code for every single review in it.

3.3 Review Extraction

The third and final step is to finally go through every review URL and scrape its contents. For that, we used the requests [5] library for Python [8] and BeautifulSoup [6] again.

After getting the raw HTML with the requests library we use BeautifulSoup to traverse the HTML tree and get the information we want such as the title, subtitle, author, date, the main review content, and the score.

Then using the CSV library in Python we create the final product of the pipeline which is the **ign.csv** file.

4 Data Characterization

We produced numerous documents for data characterization, which helped us analyze the data more easily. We created graphs using Matplotlib [1], Seaborn [2], and Pandas [7].

4.1 Number of Reviews per Year

The first graph we created presents the Number of Reviews per Year, from 1996 until today, 2024. We can see that most reviews were published between 2007 and 2014, which correlates to the time when most video games were released to the public. Many consider this the prime of video game launches.

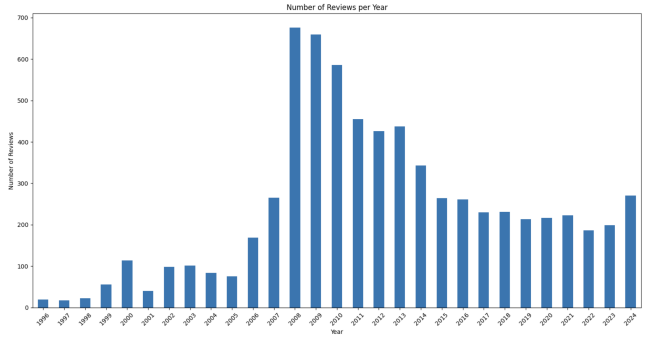

Figure 2: Reviews Per Year

4.2 Number of Reviews per Month

Figure [5] shows the number of reviews published throughout the year. We can clearly see a pattern, where the highest number of reviews publications is between the months of September and December. Normally, this is the time of the year when companies decide to launch their new video games, just before the holiday season.

4.3 Average Score per Year and Grouped Distribution of Review Scores

Figure [6] shows the score evolution over time. Back in the early two-thousands, the scores were relatively low, always in between 1 and 4. Only in 2009, we saw the first major increase in game scores. Until today, scores have not stopped increasing. We even have numerous video games maximum-rated, with a score of 10. This could be explained by the technological advancements, where developers started having the devices to produce better graphics

and more immersive experiences for the players, resulting in higher scores. Another reason could be the fact that studios nowadays have way bigger budgets, so they can hire more developers and have more polished and advanced games.

Figure [7] shows almost the same data. This time around we decided to group review scores, o clearly see the score evolution over time. We can also see the number of reviews published in those years, on top of the histograms.

4.4 Average Score by Top 20 Reviewers and Top 20 Reviewers by Number of Reviews

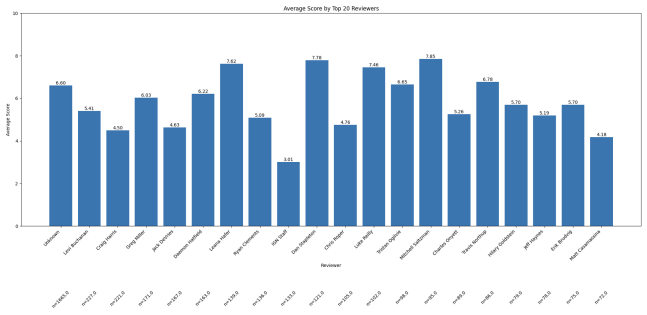

Figure 3: Average Score by Top 20 Reviewers

We also thought it would be interesting to analyze which authors tend to give higher reviews. We see that there are clearly some authors that give higher scores, maybe because some may have personal biases or preferences towards some styles or genres of games.

We also created a graph to see the author with the most published reviews (Figure [8]), and we can conclude that Levi Buchanan has 227 video games reviewed. Not far off is Craig Harris, with an astonishing 221 video games reviewed.

4.5 Word Cloud of Game Titles, Subtitles, and Content

Figure 4: Word Cloud of Game Titles, Subtitles, and Content

Word Cloud is a useful way to see which words stand out the most. In the case of the titles, we can see that the words used more often are in the first place "Review". The dataset consists of reviews, which include this term in all titles. "iPhone" because mobile games are gaining popularity, and IOS is one of the best platforms for mobile gaming. The word "Game" is fundamental and applies to all entries of the dataset.

In Figure [9], we can see that the words "Game", "New", "Fun", "Good Time", and "Best" appear regularly. This repetition shows the excitement surrounding new video game releases.

In Figure [10], we can clearly see that the word that appears the most is "Game", as the central topic of the reviews, followed by "Time", almost certainly associated with the good or bad times we can have while playing a video game, and, at last, the word "Player", that defines the person engaging with the game.

5 Search Scenarios

We identified some possible search scenarios that users would be keen to use when exploring for game reviews. These are, in our opinion, the most common ways a user would interact with our search engine:

- **Titles:** The first and most used search scenario would be to search game reviews by entering specific game titles. We think this would also be the fastest way for each user to find the review they are looking for.
- **Author:** The second possible search scenario would be to search video game reviews based on the author. It is known that some users like to find reviews from specific authors, who follow and trust their opinions.
- **Scores:** The third possible scenario would be to search for a review based on the score. A user could only be interested in games with, for example, a score above 8.
- **Subtitles:** The fourth search scenario would be to search for reviews based on their subtitles, which sometimes contain keywords, like we saw thanks to the Word Cloud analysis.
- **Date:** The fifth search scenario consists of searching for a review based on the publish date. A user with a gaming machine from the latest generation would only be interested in finding reviews for games compatible with his device, and not searching for game reviews of games some years old.
- **Genre:** The sixth and last search scenario we think would be realistic is to search for a video game review based on the genres. Normally people have a tendency to like specific types of games, so allowing the user to search by genre would be a very welcome functionality.

6 Conclusions

This first milestone was concluded with success with the preparation and characterization of the dataset selected. We have searched for datasets, chose one, performed exploratory data analysis, and prepared and documented a data processing pipeline. We have identified the conceptual model for the data domain, and defined and characterized the documents in the final collection. We are satisfied with the work we have done, and we can't wait for more opportunities to refine, both our dataset and our project as a whole.

References

[1] John D. Hunter et al. 2007. Matplotlib: A 2D Graphics Environment. https://matplotlib.org/. Accessed: 2024-10-15.
[2] Michael Waskom et al. 2012. Seaborn: Statistical Data Visualization. https://seaborn.pydata.org/. Accessed: 2024-10-15.
[3] Jason Huggins. 2004. Selenium. https://www.selenium.dev
[4] IGN Entertainment, Inc. 1996. IGN: Video Game Reviews, News, and Trailers. https://www.ign.com/. Accessed: 2024-10-15.
[5] Kenneth Reitz. 2011. Requests Library. https://pypi.org/project/requests/
[6] Leonard Richardson. 2004. BeautifulSoup. https://pypi.org/project/beautifulsoup4/
[7] The Pandas Development Team. 2008. Pandas: Python Data Analysis Library. https://pandas.pydata.org/. Accessed: 2024-10-15.
[8] Guido van Rossum. 1991. Python. https://www.python.org

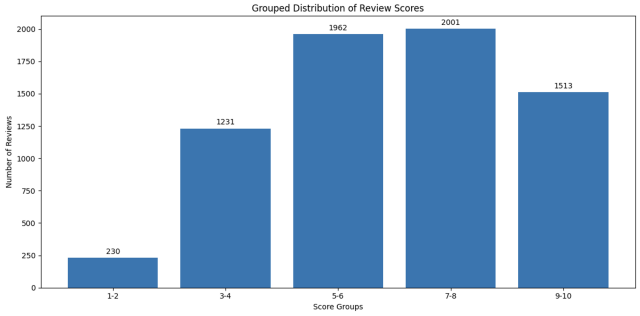

Figure 7: Distribution Review Scores

A Annexes

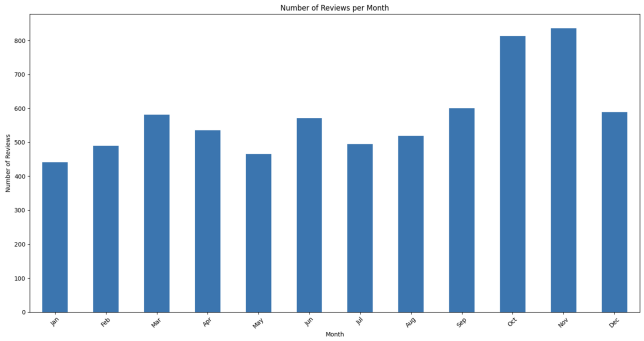

Figure 5: Reviews per Month

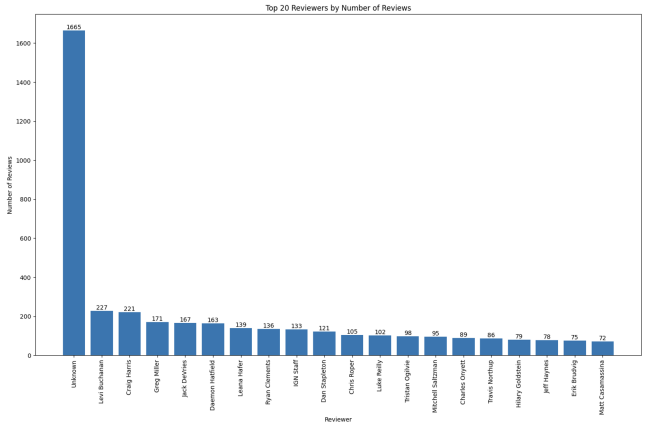

Figure 8: Top20 Reviewers

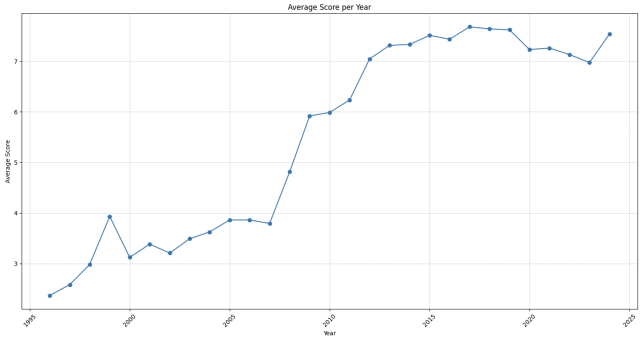

Figure 6: Scores per Year

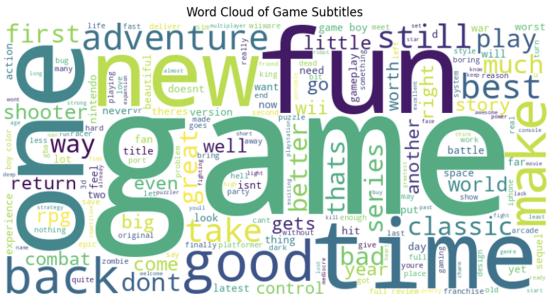

Figure 9: Word Cloud for Subtitles

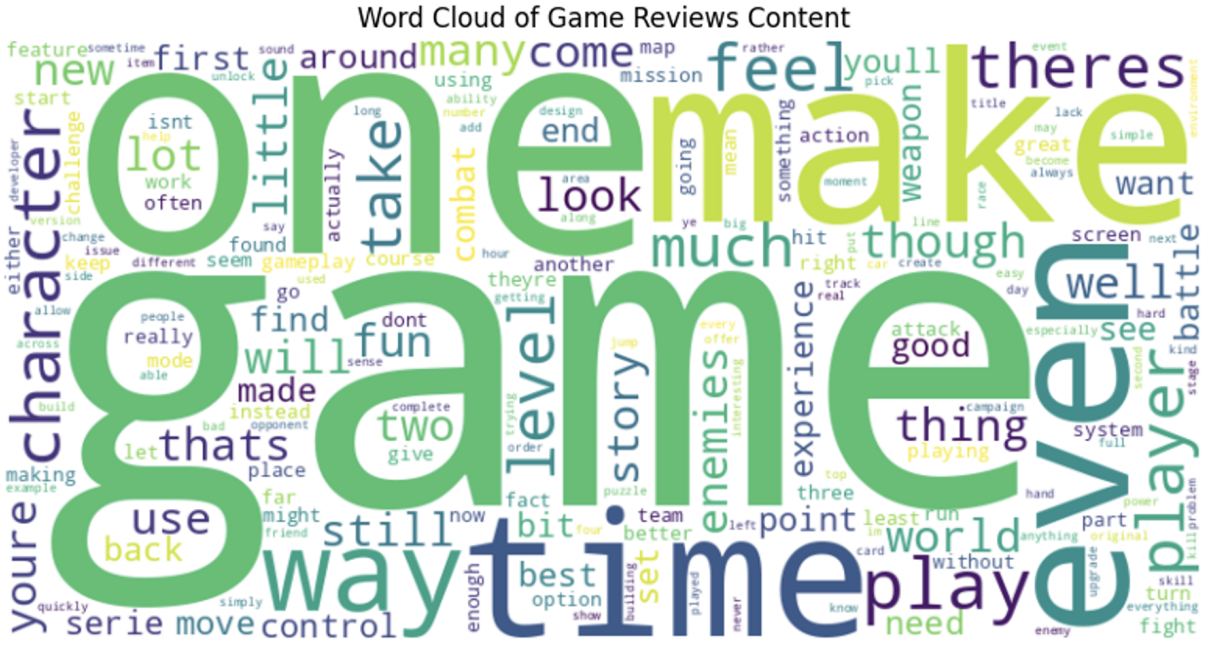

Figure 10: Word Cloud for Content

Figure 11: Data Pipeline