Processing and Retrieval of data from the Steam Library

Luís Afonso
Faculdade de Engenharia
Universidade do Porto
Porto, Porto
up201406189@edu.fe.up.pt

Nuno Oliveira
Faculdade de Engenharia
Universidade do Porto
Porto, Porto
up201806525@edu.fe.up.pt

***Abstract*—This work's goal was to create an information search system from scratch capable of handling non-trivial queries. We started by choosing and collecting data, which then needed to be merged and processed so we could analyze and better understand it. We used a dataset from Kaggle as the primary source of information which we then complemented with reviews scraped from Steam's website. Using Solr we defined an appropriate schema, indexed the fields we deemed relevant, and proceeded to provide meaningful queries in the context of our dataset. We then manually evaluated the queries results which were very acceptable. Finally, we developed a user interface for the system, added multiple new features aiming to improve the tool's performance while also providing a better user experience, and compared new results to the ones obtained before this last iteration.**

***Index Terms*—Steam, games, data collection, data cleaning, data preparation, information retrieval, search system, indexing, evaluation**

I. Introduction

Steam is a video game digital distribution service by Valve. It was launched as a standalone software client in September 2003 and as of today the biggest service of its kind with a library of over 50 thousand games. Although its search engine generally presents very good results it lacks a bit on the full-text search department and with this project, we aim to develop an information retrieval tool that serves that exact purpose.

This report is split into three major sections. The first one describes the work of preprocessing, preparing, and analyzing an existing dataset on Kaggle [1] containing information about games on Steam up to 2019. To complement and enrich the data we were working on, we added the ten most helpful reviews of each game to the dataset which were scraped directly from Steam's website. We will also explain in more detail the preparation work on the original dataset and the scraped content, the conceptual data model of the final dataset, some relevant data characterization, and possible future queries to run.

Section two is about the implementation of the information retrieval tool and the manual evaluation of queries ran. It contains detailed information about the tool choice, the resulting documents, indexes (schema), and the results for the queries, such as information needs, what was obtained, and evaluation metrics.

Section three is where we describe to user interface developed and explain all new features developed. These include synonyms, cross-language retrieval, result grouping, multiple filters, and predictive text for text boxes in the filters. To understand how the system had changed, we compared its performance after implementing all the new features to the performance in section two.

II. Data Preparation

To collect the information we deemed convenient, we decided to use a Kaggle dataset, called Steam games complete dataset, which consists of a single CSV file with data about roughly 40 thousand games, distributed over 20 columns. After combing through the dataset, we identified several problems and felt that we might not have enough textual data required for the following milestones of the project.

With that in mind, we scraped off of Steam's website [2] the ten most helpful reviews for each game in our dataset, which resulted in a large JSON file.

A. Data Cleaning and Refinement

After collecting all the data and examining it with openrefine [3] we identified multiple problems with the dataset.

- Some entries were bundles instead of games with a lot of missing information
- Characters in Mandarin which were not useful
- Multiple entries with missing values on the price column
- Games with no release date
- All entries on the publisher column were repeated
- Discount prices higher than the original price (These referred to bundles that contained the game)

To deal with these issues we created a python script to process and clean the data and leave it in the format we desired. The first four problems enunciated were easily solved by simply removing the entries where any of them were present.

The second to last required some more complicated parsing because games can have more than one developer/publisher which was separated with a comma which was a problem because some developers/publishers also had commas in their name hence the harder job of cleaning. Finally, we addressed the last problem by simply removing the respective column due to it being nonsensical.

Some extra work we did on the data was:

- Convert the 'Free' string and its variants to '0' and remove the currency character ('$') to make later data exploration simpler.
- Add a column with the id of each app which we obtained by doing some simple parsing of the 'URL' column
- Drop the 'type' column because after all the processing the only remaining type was 'app' and so its content was useless

As for the reviews we scraped, we only decided to save the textual data since it was the only aspect where our dataset needed enrichment.

On figure 1. we can observe the pipeline of this whole process, including the final data filtering and visualization scripts used for the "Data Characterization" subsection.

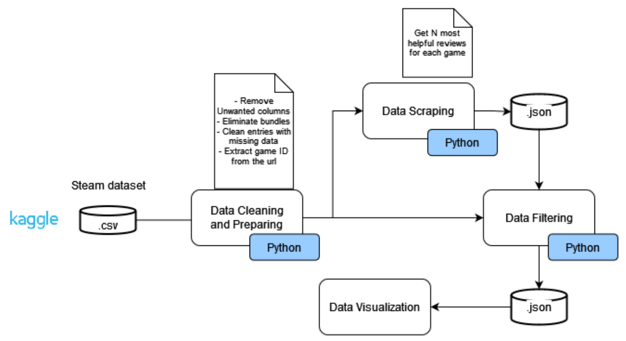

Get N most helpful reviews for each game
- Remove Unwanted columns - Eliminate bundles - Clean entries with missing data - Extract game ID from the url
Steam dataset
kaggle
.csv
Data Cleaning and Preparing
Python
Data Scraping
Python
.json
Data Filtering
Python
.json
Data Visualization

Fig. 1. Pipeline

B. Conceptual Data Model

The model obtained, which can be found in Figure 1., consists on a main class, which holds the majority of information regarding our dataset, namely:

- **url** - game url on Steam's website
- **name** - name of the game
- **description snippet** - short abstract of the game
- **description** - more detailed information about the game's content
- **recent and all reviews** - which consists of strings with simple information regarding reviews
- **date** - release date
- **achievements** - number of achievements the game has available
- **mature** - if the game has mature content, this field describes what type of mature content contains
- **description** - more detailed information about the game's content
- **price** - the cost of the game, in US dollars
- **min and recommended requirements** - minimum and recommended computer specifications by the developer

All the other classes are very simple and contain only a single, self-explanatory field.

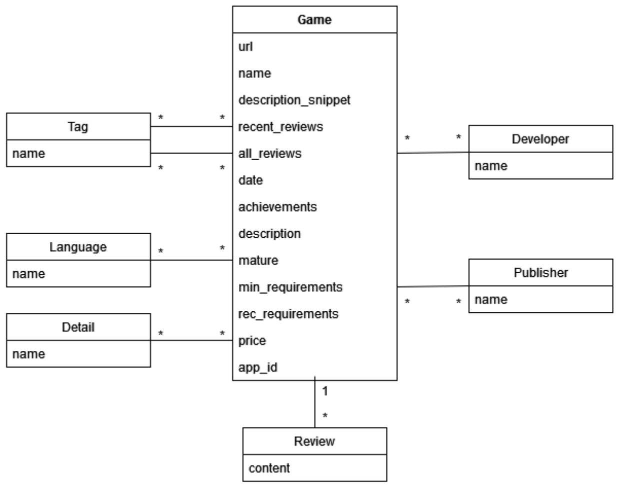


Fig. 2. Data Model

C. Dataset Characterization

The final dataset has 18 columns, including fields with only label type data, dates, some long text fields, and is accompanied by a collection of reviews written by players.

After cleaning and standardizing our dataset, we decided to collect additional information regarding our dataset to make sure it would be a good match for the tasks proposed for the remaining milestones. We also aimed to better understand what kind of content we had at our disposal, and to find interesting patterns in our data.

Regarding the distribution of game genres, Steam offers a total of 426 different tags, and to evaluate how prevalent these tags are, we decided to select 15 most common ones and bundle the rest of them in a pie graph to visually depict this information, which can be seen below in Figure 3. It is possible to see that at least one of these fifteen tags is present in 50% of the games in Steam's library.

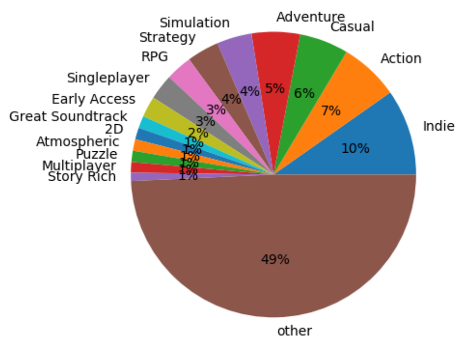


Fig. 3. Most common tags

Regarding the release dates of the games, we decided to examine how the released dates are distributed over the months, to determine if, for example, in the months leading up to the Christmas season had a higher number of games, since it would make sense, from a marketing standpoint. However, the result is quite an even dispersal as can be seen in Figure 4.

This even distribution is probably explained by the fact that most games released on Steam are not developed by big companies which tend to concentrate their releases in this period, but instead by small studios. These types of games are referred to in the industry as indie games.

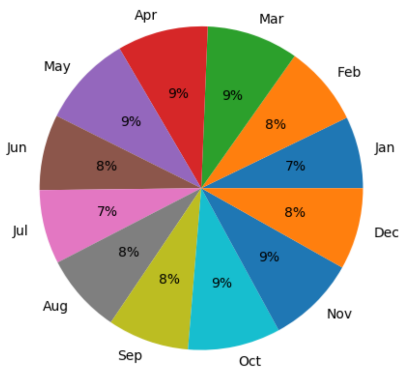


Fig. 4. Game Release distribution

We also investigated how the release dates for our games are spread throughout the years. The results obtained, visible below in Figure 5, confirm and depict the increase in the number of games published on Steam in the last few years [4], justified by the rapid growth this industry has recently experienced.

Note that the dataset only covers part of 2019, which explains the dip registered in the last year.

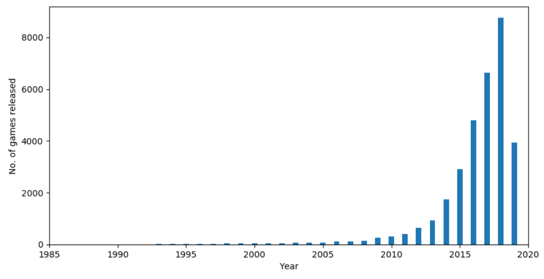


Fig. 5. Games released per year

The price of games was also subject to some exploration, to get an idea of how games are priced. Although we expected most games to be cheap, we were surprised by the number of games priced in the sub 5$ category. The following graph, Figure 6, depicts the overall prices of the games in our dataset.

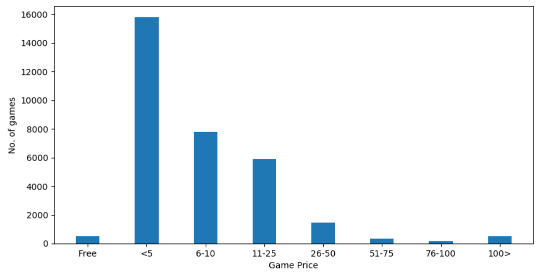


Fig. 6. Price of games at launch

Finally, we took a look at the overall feedback from the games we had collected. Considering the large number of games we had, and also the fact that a large portion of them was not developed by major companies, we were expecting a large number of them to have a negative overall score. Our results, revealed, however, a large number of games without any review(6889) and many games that did not have the necessary number of reviews to be classified in the system used by Steam(10270).

We then plotted a bar graph for the remaining games to better understand how the games were being rated by users, which can be found in Figure 7. The graph shows the contrary to what we expected to have very few games with scores lower than 'Mixed'.

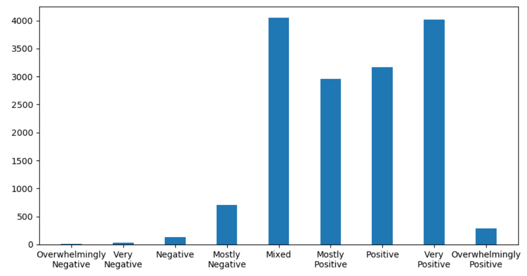


Fig. 7. User reviews

III. Information Retrieval

In this section of the report, we will cover the second milestone which aimed to implement and make use of an information retrieval tool of our choice and explore the results of FTS queries. To achieve that goal we first needed to choose a tool and produce an appropriate schema to facilitate the information retrieval process and have decent performance on our queries. We then proceeded to evaluate the performance of such queries through their precision (how many retrieved items are relevant?) and recall (how many relevant items are retrieved?). Finally, we included a brief description of some ideas we intend to explore while developing the last Milestone.

A. Tool selection

Following the recommendations given by the professors, we investigated two different tools, Apache Solr and Elasticsearch, more specifically the ELK stack [5] that combines Elasticsearch as the search and analytic engine, with Logstash being the data processing pipeline, and finally Kibana as a tool for visualizing the data. After some preliminary research we first decided to try on Elastic search since apparently it had a bigger feature repertoire and according to popular opinion was more "beginner-friendly". Unfortunately, while importing our dataset we encountered several compatibility issues which, together with time constraints, forced us to move on to Solr [6], which proved itself to be more than adequate for the proposed tasks. Solr was also much better documented than Elasticsearch, which made it simple to import our dataset, create the adequate schema, and finally query our system.

B. Documents

Our dataset resulted in a single collection with a little over 32000 documents where each document is a single game and all the information associated with it (including the reviews). We initially intended to have two separate collections, with the games and the other with only reviews, but we decided against it since multi-core searching did not seem very easy through Solr's documentation and it would substantially simplify the queries that involved reviews. We also needed to transform some fields so that they could be correctly imported to Solr, those being release_date (converted to YYYY-MM-DD), popular_tags, game_details, languages, and genre that consisted of list in form of a string with multiple values separated by ',' which needed to be converted to a proper list (a similar process was applied to merge reviews into the game document). An example of a document can be found in Appendix A.

C. Indexing

We decided to index all fields that were probable to be searched upon since our dataset will not be subject to further inserts. Taking into account that some of our fields have multiple values, for example, the list of languages, we created a field type named list_strings that employs the usage of Solr's white space tokenizer and a simple lower case filter. Regarding the fields that were rich in textual data, we devised a more complex field type capable of handling parsing the data more effectively. The field is comprised of a stop filter [7] to remove common stopwords in the English language, such as "the", "and" and the likes, followed by a Porter Stem Filter [7], which implements the Porter Stemming algorithm, a method for removing the commoner morphological and inflexional endings from words in English. An ASCII folding filter [7] was also added to remove accent and other more complex Unicode characters. Finally, an English Possessive filter [7] was used, and, as is the default for the rest of the fields, a lower case filter [7] is also present.

TABLE I
Document Custom Field Types

Field	***Tokenizer***	***Filters***
list_strings	Whitespace, Standard	Lowercase, ManagedSynonymFilterFactory
description	Classic, Standard	EnglishPossessive, PorterStem, LowerCase, ASCIIFolding ManagedSynonymFilterFactory

Table II depicts the overall types of our dataset, if they are indexed and if they are multi-valued.

TABLE II
Document Schema Fields

Field	***Type***	***Indexed***	***Multivalued***
name	text_general	Y	N
publisher	text_general	Y	Y
developer	text_general	Y	Y
desc_snippet	description	Y	N
mature_content	description	Y	N
game_description	description	Y	N
recent_reviews	plong	Y	N
all_reviews	plong	Y	N
app_id	plong	N	N
release_date	pdate	Y	N
popular_tags	list_strings	Y	Y
game_details	list_strings	Y	Y
genre	list_strings	Y	Y
languages	list_strings	Y	Y
url	text_general	N	N
minimum_requirements	text_general	N	N
recommended_requirements	text_general	N	N
achievements	pdoubles	N	N
original_price	pdoubles	Y	N

D. Retrieval Process and Evaluation

Considering the complete nature of the dataset we are working on, all reasonable and meaningful queries can be made by searching through the label-like fields, consequently, we had trouble ascertaining what type of queries could be made regarding full-text search that would be pertinent. In other words, while working on this section we concluded that our dataset doesn't require complex textual searches to retrieve the intended data because the wide range of genres and tags steam has available are often more than enough to describe games and their content. With that said we decided to try and get the same results that the normal query would achieve through the use of the game textual data, for example, instead of simply searching games with a certain tag we will search through the games textual data in search of references to that specific tag. Because we artificially enriched the dataset with reviews, full of textual data, we opted to also include them in our "textual data fields" (giving them less relative weight) along with the other relevant fields like desc_snippet, game_description, and mature_content.

Because our dataset has a lot of documents, to evaluate result precision and recall we used the first 10 and 20 games the system retrieved.

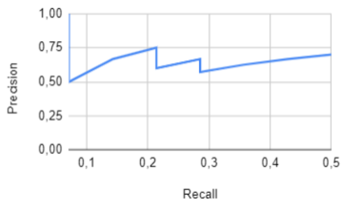
Desired result: Games similar to given game

With this search, the goal is to retrieve all games that users find similar, and to explore the query we used two games (Doom and Dark Souls). Taking into consideration that we should be exploring further into full-text search, we opted to look for mentions of the games selected in other games' descriptions and reviews from the players. This enabled us to study if players tend to mention a popular game when reviewing games similar in nature to the aforementioned videogame. As such, we queried the system for instances of the word Doom and Dark Souls in both fields and obtained the following results:

TABLE III
GAMES LIKE DOOM AND DARK SOULS

Nº	Doom	S	Dark Souls	S
1	AMID EVIL	Y	Barbarian Souls	Y
2	Solar Settlers	N	Fated Souls 2	N
3	DUSK	Y	Sekiro	Y
4	IronWolf VR	N	Fall of Light	Y
5	COMPOUND	Y	Fated Souls	N
6	Escape Goat	N	Skautfold	Y
7	Deathsmiles	Y	School Idol	N
8	Pandamonia	Y	Momodora	Y
9	The Sand Man	Y	Road Fist	N
10	Mushihimesama	Y	Left-Hand Path	Y
P@10	**0.8**		**0.6**	
Recall	**0.5**		**0.4**	

The results are interesting, although Doom gets satisfying results, Dark Souls does not. After investigating what could be the cause of this, we found out that many players tend to use this game as a comparison in regards to difficulty instead of themes. As such, games that were deemed hard were compared to Dark Souls although being nothing alike.



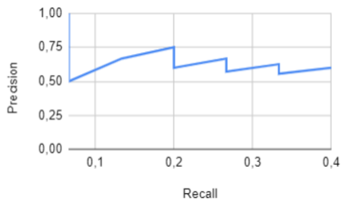


Fig. 8. P-R curves for Games Doom and Dark Souls

Desired result: Games without "Gore" content

With this search, the goal is to retrieve all games that do not contain gore content. In our dataset, some games have a short textual description of the type of mature content that can be found within the game. As such, we figured it would be interesting to query our system for a list of games that did not feature a specific kind of mature content. For this, we decided to list all occurrences that did not have the word Gore listed in the designated field.

TABLE IV
GAMES WITHOUT GORE CONTENT

Nº	Name	Success
1	Warhammer: Chaosbane	Y
2	Devil May Cry 5	N
3	TERA 3	Y
4	Life is Strange 2	Y
5	Bright Memory	N
6	Armored Battle Crew 2	Y
7	Counter-Strike: Source	Y
8	Shadow Tactics: Blades of the Shogun	N
9	Return of the Obra Dinn	Y
10	House Party	Y
P@10	**0.5**	
Recall	**0.47**	

To get fair results, we searched only through games that had the field mature content present, as it is not a mandatory field. The results are very displeasing as the precision rate is quite low. After analyzing the results, it is clear that the term gore is not always used to describe games that should be labeled as such. For example, there are multiple cases where the game mentions blood or violence. As a future improvement, synonyms could be established between these types of words, and it should vastly improve the precision of the system.

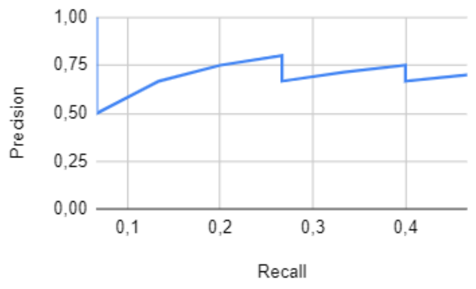


Fig. 9. P-R curves for Games With No Gore

Desired result: Games categorized as RPG (Role-playing games)

With this search, the goal is to retrieve all games classified as role-playing games. Since our dataset enables us to accurately determine the veracity of the results, even when the query returns hundreds or thousands of games, we decided to investigate if the description of the games can be used to correctly determine what kind of game it is. As such we tried to ascertain if games of the RPG type are commonly identified in either the description of the game or in the reviews of the players. For that, we queried over the game_description and reviews fields for the terms "RPG" and "Role-playing games"

TABLE V
RPG LIKE GAMES

Nº	Name	Success
1	KnightShift	Y
2	Knights of Pen and Paper	Y
3	Disgraced Role Playing Game	Y
4	Eschalon: Book III	Y
5	Wizardry 8	Y
6	Ground Under	Y
7	Realms of Arkania 1	Y
8	UnEpic	Y
9	RPG Maker VX Ace	N
10	Monster RPG 2	Y
P@10	**0.9**	
Recall	**0.5**	

Regarding the query, since we can accurately determine the number of relevant documents, and the number of relevant documents obtained, we can easily calculate both the precision and the recall of this query. For the precision, we determined that of the 5826 documents obtained only 2851 of them were RPG-like, and that the total of these types of games in our dataset numbers 12304. As such, the precision is roughly 49%, and recall sits at 47%

These values are somewhat low, but we also realize that it is not expected of a description to accurately display this type of information about the game (that's what the genre is for).

Desired result: Remastered Versions of games

For this last search, the goal is to retrieve all games that are remastered/updated versions of games that have been published in the past. In the video game industry, a remastered version is normally used to give better graphics to old games that might deserve a second release and sometimes new features, story, and gameplay. To avoid using obvious fields like that name, which usually contains the word remastered in it, we decided to look into our player reviews to see if we can predict a game as being a remastered version by the type of review they write. For this, we simply decided to look for the keyword remastered on the reviews.

TABLE VI
REMASTERED GAMES

Nº	Name	Success
1	Sniper Elite V2 Remastered	Y
2	Grim Fandango Remastered	Y
3	Full Throttle Remastered	Y
4	Voodoo Vince: Remastered	Y
5	Worms World Party Remastered	Y
6	Homeworld Remastered Collection	Y
7	Call of Duty®: Modern Warfare® Remastered	Y
8	Red Faction Guerrilla Re-Mars-tered	Y
9	Day of the Tentacle Remastered	Y
10	The Raven Remastered	Y
P@10	**1.0**	
Recall	**0.5**	

Like the previous query, in this one, we can also ascertain the accurate values regarding these two metrics. For the precision, we determined that of the total of 484 only 113 of the games fetched were RPG-like games and that the total of these types of games in our dataset numbers 714. As such the precision is roughly 23% and recall sits at 68%. The cause of this might be because users don't tend to acknowledge the new release as a remaster, nor compare it to the previous iteration. After analyzing the content of the reviews we also found that users tend to use this term when referring to the evolution of game mechanics and game engines hence the low precision.

E. Some Thoughts

Considering the mixed results obtained in the previous queries, we can conclude that it is hard to evaluate our model based on the free-text fields since players are not expected to summarize the game they have played but rather that they explain what they enjoyed or not. Similarly, developers are not supposed to post a heavy and detailed description of their game as to not spoil the fun that is discovering it by yourself.

We also ran a simple experiment comparing the results obtained when running some queries in a schemaless mode. Even tho the performance wasn't much worse at P@10, it came as no surprise to us that our model was much more flexible when dealing with inputs, for example, there were multiple instances of the word 'remaster' not being retrieved when querying for 'Remastered', which is evidence of the importance of a Porter Stemming filter like the one present in the schema system.

Although we could not accurately rate the system developed on this matter, the system could evolve into a tool to study how players write their reviews and how they could be connected to the description of the game to try to gain insightful information.

IV. Search System

In this last section before conclusions, we will go over the improvements we made to the previously developed system. Also, to make full use of the tool developed, we implemented a user interface, with some additional features, that offers a more pleasant user experience when using the information retrieval tool.

A. Revisions Introduced

Before starting working on this last milestone we revisited subsection D (Retrieval Process and Evaluation) from the Information Retrieval section. We needed to do this because we didn't have much information in terms of evaluation and evaluation methods and so, we added recall values to all queries, P-R curves, and adjusted the analysis itself.

B. User Interface

Our UI consists of 3 simple pages.

One for querying the dataset that allows us to choose the field of the document/game where we want to search and also remove a certain type of game from the search space (this is achieved through the use of synonyms and will be explained in a following subsection):

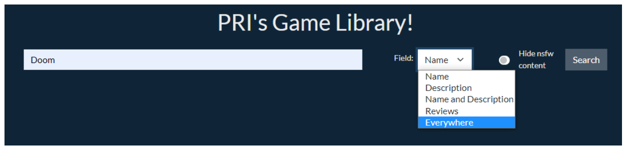
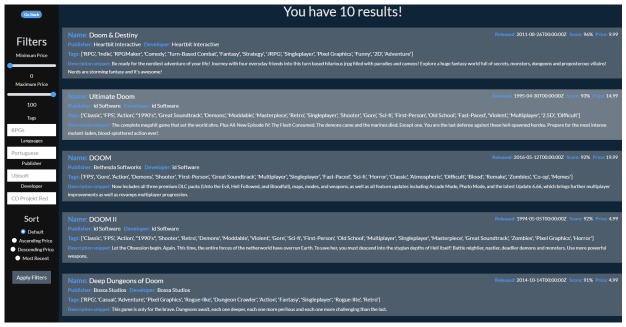


Fig. 10. Search Page

One for analyzing query results and applying a multitude of filters that allow for further specification of what type of documents we're searching for by choosing the game's price range, tags, languages, developer, publisher and even sort the results by ascend or descending price or date:



You have 10 results!

Name: Doom & Destiny Released: 2011-08-26T00:00:00Z Score: 96% Price: 9.99
Publisher: Heartbit Interactive Developer: Heartbit Interactive
Tags: ['RPG', 'Indie', 'RPGMaker', 'Comedy', 'Turn-Based Combat', 'Fantasy', 'Strategy', 'JRPG', 'Singleplayer', 'Pixel Graphics', 'Funny', '2D', 'Adventure']
Description snippet: Be ready for the nerdiest adventure of your life! Journey with four everyday friends into this turn based hilarious jrpg filled with parodies and cameos! Explore a huge fantasy world full of secrets, monsters, dungeons and preposterous villains! Nerds are storming fantasy and it's awesome!

Name: Ultimate Doom Released: 1995-04-30T00:00:00Z Score: 93% Price: 14.99
Publisher: id Software Developer: id Software
Tags: ['Classic', 'FPS', 'Action', "1990's", 'Great Soundtrack', 'Demons', 'Moddable', 'Masterpiece', 'Retro', 'Singleplayer', 'Shooter', 'Gore', 'Sci-fi', 'First-Person', 'Old School', 'Fast-Paced', 'Violent', 'Multiplayer', '2.5D', 'Difficult']
Description snippet: The complete megahit game that set the world afire. Plus All-New Episode IV: Thy Flesh Consumed. The demons came and the marines died. Except one. You are the last defense against these hell-spawned hordes. Prepare for the most intense mutant-laden, blood-splattered action ever!

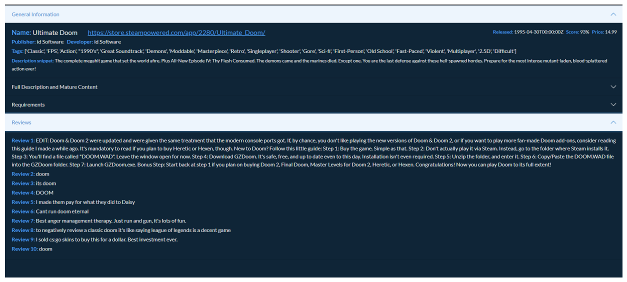
Name: DOOM Released: 2016-05-12T00:00:00Z Score: 92% Price: 19.99
Publisher: Bethesda Softworks Developer: id Software
Tags: ['FPS', 'Gore', 'Action', 'Demons', 'Shooter', 'First-Person', 'Great Soundtrack', 'Multiplayer', 'Singleplayer', 'Fast-Paced', 'Sci-fi', 'Horror', 'Classic', 'Atmospheric', 'Difficult', 'Blood', 'Remake', 'Zombies', 'Co-op', 'Memes']
Description snippet: Now includes all three premium DLC packs (Unto the Evil, Hell Followed, and Bloodfall), maps, modes, and weapons, as well as all feature updates including Arcade Mode, Photo Mode, and the latest Update 6.66, which brings further multiplayer improvements as well as revamps multiplayer progression.

Name: DOOM II Released: 1994-05-05T00:00:00Z Score: 92% Price: 4.99
Publisher: id Software Developer: id Software
Tags: ['Classic', 'FPS', 'Action', "1990's", 'Shooter', 'Retro', 'Demons', 'Moddable', 'Violent', 'Gore', 'Sci-fi', 'First-Person', 'Old School', 'Multiplayer', 'Singleplayer', 'Masterpiece', 'Great Soundtrack', 'Zombies', 'Pixel Graphics', 'Horror']
Description snippet: Let the Obsession begin. Again. This time, the entire forces of the netherworld have overrun Earth. To save her, you must descend into the stygian depths of Hell itself! Battle mightier, nastier, deadlier demons and monsters. Use more powerful weapons.

Name: Deep Dungeons of Doom Released: 2014-10-14T00:00:00Z Score: 91% Price: 4.99
Publisher: Bossa Studios Developer: Bossa Studios
Tags: ['RPG', 'Casual', 'Adventure', 'Pixel Graphics', 'Rogue-like', 'Dungeon Crawler', 'Action', 'Fantasy', 'Singleplayer', 'Rogue-lite', 'Retro']
Description snippet: This game is only for the brave. Dungeons await, each one deeper, each one more perilous and each one more challenging than the last.

Fig. 11. Results

One to fully visualize a specific document/game and also contains the link to the game's webpage on Steam (we decided to have this page because each game has a lot of textual information that would make visualizing results harder if it were all shown together):



Name: Ultimate Doom https://store.steampowered.com/app/2280/Ultimate_Doom/ Released: 1995-04-30T00:00:00Z Score: 93% Price: 14.99
Publisher: id Software Developer: id Software
Tags: ['Classic', 'FPS', 'Action', "1990's", 'Great Soundtrack', 'Demons', 'Moddable', 'Masterpiece', 'Retro', 'Singleplayer', 'Shooter', 'Gore', 'Sci-fi', 'First-Person', 'Old School', 'Fast-Paced', 'Violent', 'Multiplayer', '2.5D', 'Difficult']
Description snippet: The complete megahit game that set the world afire. Plus All-New Episode IV: Thy Flesh Consumed. The demons came and the marines died. Except one. You are the last defense against these hell-spawned hordes. Prepare for the most intense mutant-laden, blood-splattered action ever!

Review 1: EDIT: Doom & Doom 2 were updated and were given the same treatment that the modern console ports got. If, by chance, you don't like playing the new versions of Doom & Doom 2, or if you want to play more fan-made Doom add-ons, consider reading this guide I made a while ago. It's mandatory to read if you plan to buy Heretic or Hexen, though. New to Doom? Follow this little guide: Step 1: Buy the game. Simple as that. Step 2: Don't actually play it via Steam. Instead, go to the folder where Steam installs it. Step 3: You'll find a file called "DOOM.WAD". Leave the window open for now. Step 4: Download GZDoom. It's safe, free, and up to date even to this day. Installation isn't even required. Step 5: Unzip the folder, and enter it. Step 6: Copy/Paste the DOOM.WAD file into the GZDoom folder. Step 7: Launch GZDoom.exe. Bonus Step: Start back at step 1 if you plan on buying Doom 2, Final Doom, Master Levels for Doom 2, Heretic, or Hexen. Congratulations! Now you can play Doom to its full extent!
Review 2: doom
Review 3: its doom
Review 4: DOOM
Review 5: I made them pay for what they did to Daisy
Review 6: Cant run doom eternal
Review 7: Best anger management therapy. Just run and gun, it's lots of fun.
Review 8: to negatively review a classic doom it's like saying league of legends is a decent game
Review 9: I sold cs:go skins to buy this for a dollar. Best investment ever.
Review 10: doom

Fig. 12. Game Page

C. System Improvements

As previously stated we introduced several improvements to our information retrieval tool.

The first one was the addition of synonyms. Our dataset benefits a lot from a well-made list of synonyms because a single genre can be described with various terms that generally have the same meaning. With synonyms, our queries can now retrieve relevant games that wouldn't otherwise be considered as such (in the next subsection we will demonstrate this).

The second improvement, already mentioned, was the addition of filters to the result page. With filters applied, we new, more specific query will be generated and sent to Solr. Since our filters include text boxes for fields with label-like information we also added predictive text to those input fields to provide a better and smoother user experience.

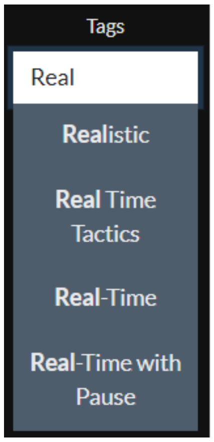
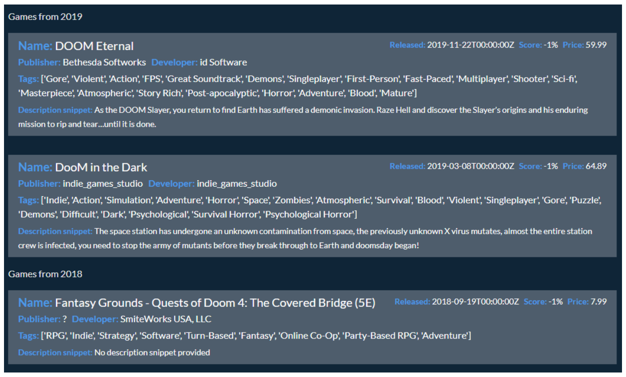


Fig. 13. Predictive text in filters

The third improvement was implementing cross-language retrieval. When a search is made, before creating the query and sending it to Solr, we first auto-detect text language and translate it to English (if the detected language isn't already English). This feature also greatly improves our tool's performance as will be seen in the next section.

For the final improvement, we decided to add a simple grouping of games retrieved from the query. For example, if the user decides to sort the query result by most recent, each game will be under a section corresponding to the year of its release date.



Games from 2019

Name: DOOM Eternal Released: 2019-11-22T00:00:00Z Score: -1% Price: 59.99
Publisher: Bethesda Softworks Developer: id Software
Tags: ['Gore', 'Violent', 'Action', 'FPS', 'Great Soundtrack', 'Demons', 'Singleplayer', 'First-Person', 'Fast-Paced', 'Multiplayer', 'Shooter', 'Sci-fi', 'Masterpiece', 'Atmospheric', 'Story Rich', 'Post-apocalyptic', 'Horror', 'Adventure', 'Blood', 'Mature']
Description snippet: As the DOOM Slayer, you return to find Earth has suffered a demonic invasion. Raze Hell and discover the Slayer's origins and his enduring mission to rip and tear...until it is done.

Name: DooM in the Dark Released: 2019-03-08T00:00:00Z Score: -1% Price: 64.89
Publisher: indie_games_studio Developer: indie_games_studio
Tags: ['Indie', 'Action', 'Simulation', 'Adventure', 'Horror', 'Space', 'Zombies', 'Atmospheric', 'Survival', 'Blood', 'Violent', 'Singleplayer', 'Gore', 'Puzzle', 'Demons', 'Difficult', 'Dark', 'Psychological', 'Survival Horror', 'Psychological Horror']
Description snippet: The space station has undergone an unknown contamination from space, the previously unknown X virus mutates, almost the entire station crew is infected, you need to stop the army of mutants before they break through to Earth and doomsday began!

Games from 2018

Name: Fantasy Grounds - Quests of Doom 4: The Covered Bridge (5E) Released: 2018-09-19T00:00:00Z Score: -1% Price: 7.99
Publisher: ? Developer: SmiteWorks USA, LLC
Tags: ['RPG', 'Indie', 'Strategy', 'Software', 'Turn-Based', 'Fantasy', 'Online Co-Op', 'Party-Based RPG', 'Adventure']
Description snippet: No description snippet provided

Fig. 14. Results grouped by date

D. *Performance comparison*

To more easily test the impact synonyms have on our system, and ultimately because we thought it was a good feature, the search page contains a button to hide results that contain improper results. This was achieved by making use of synonyms, matching the keyword "NSFW" to multiple other words or expressions used to describe content not appropriate for children.

The effect of this feature (synonyms) can be ascertained by comparing the results obtained previously in the second query with the ones obtained in the new system.

TABLE VII
GAMES WITHOUT GORE CONTENT - WITH SYNONYMS

Nº	Name	Success
1	BATTLETECH	Y
2	Human: Fall Flat	Y
3	They Are Billions	Y
4	Warhammer: Chaosbane	Y
5	For The King	Y
6	Danganronpa V3	Y
7	Stonehearth	Y
8	Team Sonic Racing	Y
9	Titan Quest Anniversary Edition	Y
10	Portal	Y
P@10	**1**	
Recall	**0.5**	

Even though recall barely improved (47% to 50%), the precision of the first 10 results has doubled (now sitting at 100%). It is clear to see how synonyms play a big role in querying our dataset.

Since we also made our system capable of handling queries in multiple languages, here's what the tool finds as relevant when we search, in Portuguese, "Jogos de tiros":

TABLE VIII
SHOOTING GAMES - PORTUGUESE

Nº	Name	Success
1	Blue Revolver	Y
2	Deathsmiles	Y
3	Triana's Project: Battle Splash 2.0	Y
4	Mushihimesama	Y
5	DoDonPachi Resurrection	Y
6	Fantastic Danmaku Festival Part II	Y
7	Murasaki	Y
8	Super Amazing Wagon Adventure	Y
9	Heroine of the Sniper	Y
10	DEADBOLT	Y
P@10	**1**	
Recall	**0.5**	

Before supporting cross-language retrieval, this search would not find any relevant games and now, it has precision at 10 of 100%. While not directly improving system performance, because the user could have just manually translated the text before querying we deem this as a very important feature due to the better user experience and inclusion it provides.

V. CONCLUSIONS

This work enabled us to experiment with a previously crafted dataset and also the opportunity to explore other methods of information collection such as scraping, even if we did not delve too deep into it. We encountered several roadblocks while refining the dataset, all of them which we were able to overcome. This gave us a better understanding of the tools and methods available to gather large amounts of data, and how to better process it so that it can be worked on to provide meaningful data ready to be explored.

Through data analysis and while working on the second milestone we concluded that our dataset was not a great one to full-text search on since the games have a huge variety of tags and genres (label-like fields) to describe them and these are what is generally used to filter the relevant videogames. Despite that, we still tried to explore relevant and meaningful queries to the theme and came to acceptable results.

The third and final milestone was very satisfactory to work on because we got to make our tool usable and added a few quality-of-life features to provide a smooth experience while using it. It was also very enjoyable to see how the system had improved when redoing the queries made in the second milestone and getting overall way better results.

REFERENCES

[1] Steam games complete dataset. Accessed on: October 29th, 2021. [Online] Available: https://www.kaggle.com/trolukovich/steam-games-complete-dataset
[2] Steamworks documentation. Accessed on: October 29th, 2021. [Online] Available: https://partner.steamgames.com/doc/store/getreviews
[3] A free, open source, powerful tool for working with messy data. Accessed on: October 29th, 2021. [Online] Available: https://openrefine.org/
[4] Number of games released on steam worldwide from 2004 to 2021. Accessed on: November 13th, 2021. [Online] Available: https://www.statista.com/statistics/552623/number-games-released-steam/

[5] What is the ELK Stack? Accessed on: November 13th, 2021. [Online] Available: https://www.elastic.co/pt/what-is/elk-stack
[6] Apache Solr. Accessed on: December 10th, 2021. [Online] Available: https://solr.apache.org/
[7] Solr Filter Descriptions. Accessed on: December 10th, 2021. [Online] Available: https://solr.apache.org/guide/6_6/filter-descriptions.html
[8] Filme Ratings. Accessed on: January 15th, 2022. [Online] Available: https://www.motionpictures.org/film-ratings/