Datasheet for dataset “Game Reviews RPS in Recent Years”

Author: Irene Jiaying Xu

Organisation: Creative Computing Institute, University of the Arts London

Date: 06/12/2023

Datasheet template is taken from: https://github.com/fau-masters-collected-works-cgarbin/datasheet-for-dataset-template

Questions from the Datasheets for Datasets paper, v7.

Motivation

The questions in this section are primarily intended to encourage dataset creators to clearly articulate their reasons for creating the dataset and to promote transparency about funding interests.

For what purpose was the dataset created?

This dataset is created solely for learning purposes, it is part of the submitted work of the MSc DSAI Term 1 NLP course project. The goal is to extract game reviews data from https://www.rockpapershotgun.com/reviews and conduct basic text analysis to obtain insights into the reviews on Rock Paper Shotgun.

Who created the dataset (e.g., which team, research group) and on behalf of which entity (e.g., company, institution, organization)?

Irene Xu, an MSc student at UAL Creative Computing Institute.

Who funded the creation of the dataset?

The creation of this dataset is not funded.

Any other comments?

No.

Composition

Most of these questions are intended to provide dataset consumers with the information they need to make informed decisions about using the dataset for specific tasks. The answers to some of these questions reveal information about compliance with the EU's General Data Protection Regulation (GDPR) or comparable regulations in other jurisdictions.

What do the instances that comprise the dataset represent (e.g., documents, photos, people, countries)?

Each instance represents a published game review article with relevant game information (incl. page titles, URLs, game titles, developers, topic labels, briefs, main review text, and published dates) on rockpapershotgun.com, stored as individual text files as well as JSON and CSV file.

How many instances are there in total (of each type, if appropriate)?

540

Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set?

This dataset is a subset of a larger set, it contains only partial review articles (published from 1/6/2020 to 1/12/2023) from all review articles on rockpapershotgun.com.

What data does each instance consist of?

Each instance consists of the article title, URL, game title, developers, labels, brief, main review paragraphs, and published date.

Is there a label or target associated with each instance?

Labels associated with each instance were included in the ‘Labels’ column in the CSV file.

Is any information missing from individual instances?

Some instances lack information in the ‘brief’ field, which is indicated by ‘n/a’.

Are relationships between individual instances made explicit (e.g., users’ movie ratings, social network links)?

n/a

Are there recommended data splits (e.g., training, development/validation, testing)?

n/a

Are there any errors, sources of noise, or redundancies in the dataset?

The game title of each instance may not be very accurate, considering it is directly extracted from the page title. There might be situations where different game titles actually refer to the same game.

There may also be errors and incorrect data in the extracted dataset, as the HTML format behind each review page is inconsistent (especially for review articles published before 2021). The code written to extract data from a total of 11 different HTML format variations might still be insufficient, resulting in inaccurate data being extracted from the websites.

Is the dataset self-contained, or does it link to or otherwise rely on external resources (e.g., websites, tweets, other datasets)?

The dataset created in this project is self-contained. However, there is no guarantee that the source of each review article on Rock Paper Shotgun will exist, and remain constant over time.

Does the dataset contain data that might be considered confidential (e.g., data that is protected by legal privilege or by doctor-patient confidentiality, data that includes the content of individuals' non-public communications)?

No.

Does the dataset contain data that, if

viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety?

No. However, each review article is written by an individual and may be subjective. It is not recommended to consider the dataset as official or decisive material.

Does the dataset relate to people?

Some of the instances might contain the names of the writers.

Does the dataset identify any subpopulations (e.g., by age, gender)?

It is possible to identify the gender of some of the writers as some writers' names are mentioned in the article which is written in the third person (with the use of he or she).

Is it possible to identify individuals (i.e., one or more natural persons), either directly or indirectly (i.e., in combination with other data) from the dataset?

It is possible to identify individuals as some of the writers' names were mentioned in the article extracted from the website.

Does the dataset contain data that might be considered sensitive in any

way (e.g., data that reveals racial or ethnic origins, sexual orientations, religious beliefs, political opinions or union memberships, or locations; financial or health data; biometric or genetic data; forms of government identification, such as social security numbers; criminal history)?

It is possible to identify individuals' names and sometimes even gender, but it would be rather hard to identify other sensitive information associated with the individual, as the name of the author of each review was not included in the data extracted from the website (unless they mentioned themselves in the article).

Any other comments?

No.

Collection process

The answers to questions here may provide information that allow others to reconstruct the dataset without access to it.

How was the data associated with each instance acquired?

The data is scraped and extracted from RockPaperShotgun, a British video game journalism website. Each document/row in the dataset represent a published review article on rockpapershotgun.com, with Title, URL, Developers, Labels, Brief, and main Reviews available.

Each instance is stored as its own text file in 'game_reviews_rps_in_recent_years' folder.

What mechanisms or procedures were used to collect the data (e.g., hardware apparatus or sensor, manual human curation, software program, software API)?

BeautifulSoup4, Requests, and Loops.

If the dataset is a sample from a larger set, what was the sampling strategy (e.g., deterministic, probabilistic with specific sampling probabilities)?

This dataset is a subset of the review articles on https://www.rockpapershotgun.com/reviews. It only contains 540 review articles, articles published before 01/06/2020 were removed as most of them lack details of brief and developer information. There also isn't enough time to adjust code to accurately extract data from all review articles.

Who was involved in the data collection process (e.g., students, crowdworkers, contractors) and how were they

compensated (e.g., how much were crowdworkers paid)?

Irene Xu was involved in the data collection and web scraping process. She is a student studying MSc DSAT at UAL CCI, and this project is a school project for the course Natural Language Processing for the Creative Industries.

Over what timeframe was the data collected?

The data was collected (last updated) on 01/12/2023

Were any ethical review processes conducted (e.g., by an institutional review board)?

No.

Does the dataset relate to people?

All authors' names can be publicly viewed and found on the source website, Some of the authors' names were mentioned in the review part of this dataset.

Did you collect the data from the individuals in question directly, or obtain it via third parties or other sources (e.g., websites)?

The data is extracted from rockpapershotgun.com.

Were the individuals in question notified about the data collection?

n/a

Did the individuals in question consent to the collection and use of their data?

n/a

If consent was obtained, were the consenting individuals provided with a mechanism to revoke their consent in the future or for certain uses?

n/a

Has an analysis of the potential impact of the dataset and its use on data subjects (e.g., a data protection impact analysis) been conducted?

n/a

Any other comments?

No.

Preprocessing/cleaning/labeling

The questions in this section are intended to provide dataset consumers with the information they need to determine whether the "raw" data has been processed in ways that are compatible with their chosen tasks. For example, text that has been converted into a "bag-of-words" is not suitable for tasks involving word order.

Was any preprocessing/cleaning/labeling of the data done (e.g., discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of missing values)?

Preprocessing and cleaning were conducted at the same time when extracting data from the web pages using bs4 and regex. Detailed code and comments can be found in Python notebook '1b_web_scrape_rps_links.ipynb' and '1c_web_scrape_rps_content.ipynb'. No manual data processing is required during the process.

Was the "raw" data saved in addition to the preprocessed/cleaned/labeled data (e.g., to support unanticipated future uses)?

The raw data extracted has been saved to the ‘test output’ folder for comparison. However, it is important to note that the code used for extracting raw data is not able to accurately capture all information, leading to omissions and misinformation. Consequently, the raw data is not suitable for processing or analysing work without the removal of the incorrect data.

Is the software used to preprocess/clean/label the instances available?

The code used to preprocess/cleaning/label can be found in the submitted zip file and on GitHub. The link is included in the README.md.

Any other comments?

No.

Uses

These questions are intended to encourage dataset creators to reflect on the tasks for which the dataset should and should not be used. By explicitly highlighting these tasks, dataset creators can help dataset consumers to make informed decisions, thereby avoiding potential risks or harms.

Has the dataset been used for any tasks already?

For learning purposes only - e.g., applying topic models as part of the project work. Relevant code and report can be found in the repository.

Is there a repository that links to any or all papers or systems that use the dataset?

n/a

What (other) tasks could the dataset be used for?

n/a

Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses?

The code used to extract the current dataset from rockpapershotgun.com is valid only for a specific time range (2020/6/1 - 2023/12/1). Reviews published before or after this timeframe might have a completely different HTML code and cannot be extracted or labelled using the same code.

Are there tasks for which the dataset should not be used?

This dataset is created for learning purposes only.

Any other comments?

No.

Distribution

Will the dataset be distributed to third parties outside of the entity (e.g., company, institution, organization) on behalf of which the dataset was created?

This dataset is created solely for learning purposes and will only be shared internally for grading.

How will the dataset will be distributed (e.g., tarball on website, API, GitHub)?

Internal Submission Portal, GitHub.

When will the dataset be distributed?

Decmber 2023

Will the dataset be distributed under a copyright or other intellectual property (IP) license, and/or under applicable

terms of use (ToU)?

No.

Have any third parties imposed IP-based or other restrictions on the data associated with the instances?

No.

Do any export controls or other regulatory restrictions apply to the dataset or to individual instances?

No.

Any other comments?

No.

Maintenance

These questions are intended to encourage dataset creators to plan for dataset maintenance and communicate this plan with dataset consumers.

Who is supporting/hosting/maintaining the dataset?

Until the completion and submission of the project, Irene Xu will be responsible of all supporting, hosting, and maintaining work related to the dataset.

How can the owner/curator/manager of the dataset be contacted (e.g., email address)?

Please reach out via University Email or Slack.

Is there an erratum?

No.

Will the dataset be updated (e.g., to correct labeling errors, add new instances, delete instances)?

This dataset was created as part of a student project work of the DSAI NLP course for learning purposes only. Errors (if any) will be corrected, but generally, there will not be major updates or maintenance happening after the completion and submission of the project.

If the dataset relates to people, are there applicable limits on the retention of the data associated with the instances (e.g., were individuals in question told that their data would be retained for a fixed period of time and then deleted)?

No.

Will older versions of the dataset continue to be supported/hosted/maintained?

The older versions will be stored on GitHub.

If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so?

This project and dataset will only be shared within the University Git Network. Anyone outside the network has no access and thus cannot extend/augment/build on/contribute to the dataset.

Any other comments?

No.