

# INFORME PRA<sub>1</sub>

## WEB SCRAPING

Carlos Humberto Carreño Díaz

David Barrera Montesdeoca

Tipología y ciclo de vida de los datos

23 octubre de 2020

## Contenido

ENUNCIADO .....	2
DESARROLLO .....	3
1. CONTEXTO.....	3
2. TÍTULO.....	3
3. DESCRIPCIÓN.....	3
4. REPRESENTACIÓN GRÁFICA.....	3
5. CONTENIDO.....	5
6. AGRADECIMIENTOS.....	6
7. INSPIRACIÓN.....	6
8. LICENCIA.....	7
9. CÓDIGO.....	7
10. DATASET.....	8

## Enunciado

En el presente informe se documentan los siguientes aspectos:

1. Contexto. Explicar en qué contexto se ha recolectado la información. Explique por qué el sitio web elegido proporciona dicha información.
2. Definir un título para el dataset. Elegir un título que sea descriptivo.
3. Descripción del dataset. Desarrollar una descripción breve del conjunto de datos que se ha extraído (es necesario que esta descripción tenga sentido con el título elegido).
4. Representación gráfica. Presentar una imagen o esquema que identifique el dataset visualmente
5. Contenido. Explicar los campos que incluye el dataset, el periodo de tiempo de los datos y cómo se ha recogido.
6. Agradecimientos. Presentar al propietario del conjunto de datos. Es necesario incluir citas de investigación o análisis anteriores (si los hay).
7. Inspiración. Explique por qué es interesante este conjunto de datos y qué preguntas se pretenden responder.
8. Licencia. Seleccione una de estas licencias para su dataset y explique el motivo de su selección:
  - Released Under CCo: Public Domain License
  - Released Under CC BY-NC-SA 4.0 License
  - Released Under CC BY-SA 4.0 License
  - Database released under Open Database License, individual contents under Database Contents License
  - Other (specified above)
  - Unknown License
9. Código. Adjuntar el código con el que se ha generado el dataset, preferiblemente en Python o, alternativamente, en R.
10. Dataset. Publicación del dataset en formato CSV en Zenodo (obtención del DOI) con una breve descripción.

# Desarrollo

## 1. CONTEXTO.

En los últimos años, la proliferación de usuarios en la industria de videojuegos ha crecido de manera constante. Esto ha generado grandes volúmenes de datos para los cuales las diferentes compañías del sector han incluido diferentes técnicas de análisis para mejorar sus productos, tomar decisiones de mercadeo, elegir las promociones con más probabilidades de tener éxito e incluso seleccionar el género y los temas de los próximos juegos.

Los comentarios de los jugadores, por ejemplo, representan un valioso recurso que permite conocer no sólo la impresión que tienen de los juegos, sino ideas que permiten mejorar la experiencia de usuario y establecer fuentes de ideas para el desarrollo nuevas funcionalidades.

Para la muestra, se recopiló información de las valoraciones del juego Fall Guys, el cual ha sido tendencia en los últimos meses, no sólo por la jugabilidad, sino por la creciente demanda de juegos de competencia masiva. El sitio web elegido, [www.steamcommunity.com](http://www.steamcommunity.com), permite que los usuarios que han adquirido el juego puedan subir reseñas sobre el juego indicando, entre otra información, el número de horas que han utilizado en jugarlo.

## 2. TÍTULO.

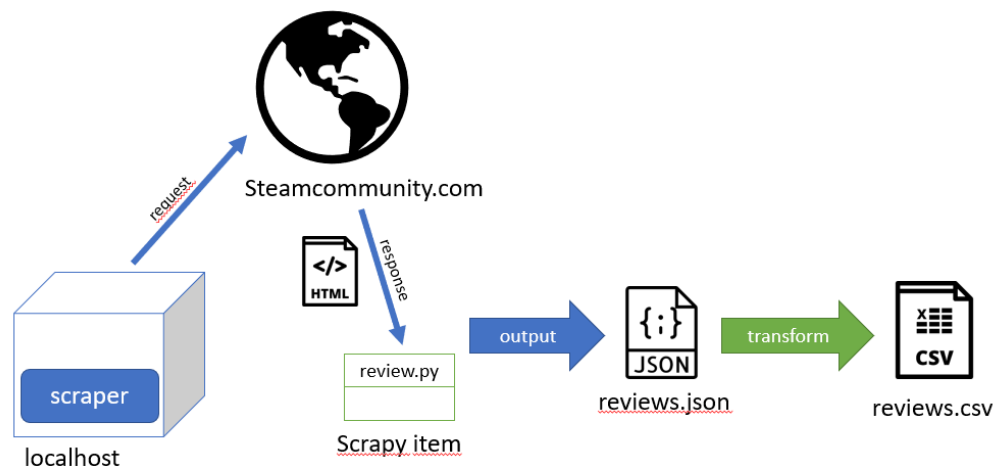
El título del dataset recolectado es: Reseñas de usuarios sobre el juego “*Fall Guys: Ultimate Knockout*” en idioma inglés.

## 3. DESCRIPCIÓN.

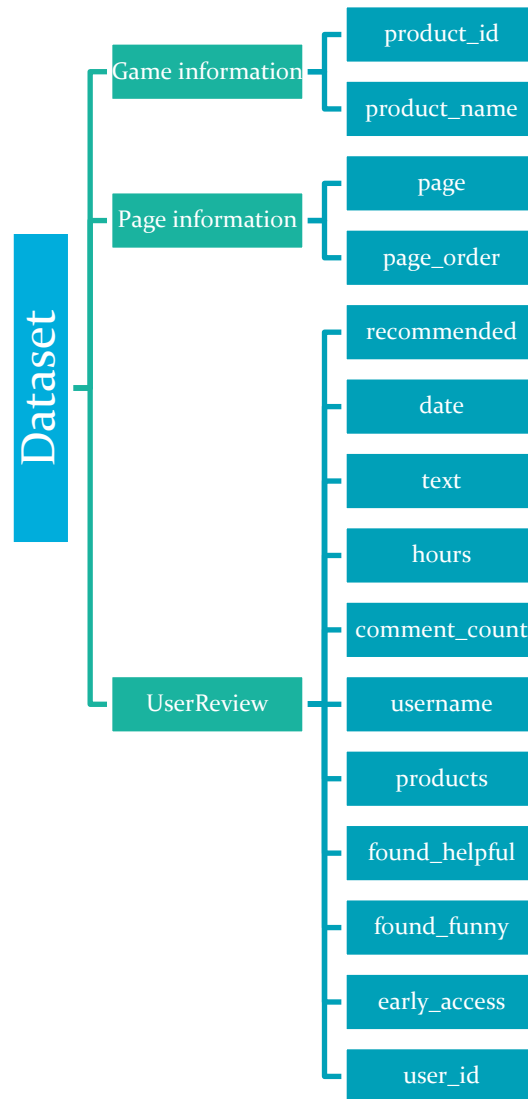
El dataset presenta información asociada a las reseñas de los usuarios del juego “Fall Guys: Ultimate Knockout”, indicando comentarios y las recomendaciones valoraciones de la reseña por parte de otros usuarios en idioma inglés.

## 4. REPRESENTACIÓN GRÁFICA.

El proyecto utiliza el siguiente flujo de datos, partiendo desde el equipo local y utilizando el scraper para acceder al sitio seleccionado, transformando cada respuesta en un objeto que podrá ser almacenado en un archivo json que posteriormente es transformado en csv.



La estructura general del dataset, se presenta a continuación:



## 5. CONTENIDO.

El scraper consulta la página de reseñas del juego. Dado que la página no carga todas las reseñas de manera inmediata, se tuvo que realizar una serie de callbacks en el script para que cargue las reseñas de la siguiente página y continuar hasta que recolecte todos los reviews en el idioma especificado. Para la recolección de los datos se utilizó web scrapping, la recolección duró aproximadamente **2 horas y 35 minutos** y se extrajeron un total de **105.247** registros. El periodo de tiempo usado va desde el lanzamiento del juego en **agosto de 2020 a noviembre de 2020**, recibiendo una gran cantidad de registros para un periodo corto de tiempo.

La descripción de los campos que contiene el dataset se detalla a continuación:

Información del juego:

product\_id: identificador único del juego

product\_name: nombre del juego

Información de la página:

page: página en la que está la reseña

page\_order: orden de la reseña en la página

Información de la reseña:

recommended: booleano, indica si el usuario recomienda o no el juego

date: fecha de realización del comentario

text: detalles del comentario

hours: horas que ha jugado el juego comentado

found\_helpful: cantidad de usuarios que encontraron útil la reseña

found\_unhelpful: cantidad de usuarios que NO encontraron útil la reseña

found\_funny: cantidad de usuarios que encontraron divertido la reseña

comment\_count: cantidad de comentarios de otros usuarios en la reseña

username: nombre de usuario de quien hace la reseña

user\_id: identificador único de usuario de quien hace la reseña

products: cantidad de productos adquiridos por el usuario

early\_access: booleano, indica si el usuario adquirió el producto mediante acceso adelantado (preventa)

## 6. AGRADECIMIENTOS.

Agradecimientos al sitio web <https://steamcommunity.com/> propietaria de los datos, los cuales son recopilados gracias al uso de los usuarios de la plataforma Steam, la cual es una plataforma de distribución digital de videojuegos desarrollada por Valve Corporation.

El sitio de los datos posee un archivo robots.txt con la siguiente configuración, facilitando el proceso de extracción:

```
User-agent: *
Disallow: /actions/
Disallow: /linkfilter/
Disallow: /tradeoffer/
Disallow: /trade/
Disallow: /email/
Host: steamcommunity.com
```

Adicionalmente, se hace el reconocimiento que el interés en la generación de este dataset surge como respuesta al trabajo realizado en análisis de sentimientos en juegos realizado en investigaciones, principalmente las que se presentan a continuación:

- Bais, R., & Odek, P. (2017). Sentiment Classification on Steam Reviews.
- Zuo, Z. (2018). Sentiment Analysis of Steam Review Datasets using Naive Bayes and Decision Tree Classifier.

## 7. INSPIRACIÓN.

El análisis de sentimientos se centra en realizar minería de opiniones mediante el procesamiento del lenguaje natural y la minería de textos. La industria de los videojuegos es una industria de alta competencia que requiere conocer a los usuarios, sus opiniones e intereses, así como las tendencias del mercado y la demanda. Como muestra se tomó como referencia el fenómeno generado por juegos de tipo “Battle Royale” (batalla real), en específico el juego “Fall Guys” el cual cuenta con más de un

millón de seguidores en Twitter, puesto número uno en la plataforma Twitch y que ha generado un fenómeno cultural alrededor de ello, en general las relaciones entre competidores <sup>1 2</sup>.

Gracias a esto, es posible realizar la recopilación y preparación de datos hasta la clasificación final mediante diferentes modelos y clasificadores de selección de características que permitan a las empresas a aumentar las ganancias potenciales en el mercado global de productos digitales, a partir de la identificación de los intereses de los usuarios.

Mediante el uso de las reseñas (se seleccionó el idioma inglés ya que provee el conjunto de datos más grande posible), se puede realizar un análisis de sentimientos de los comentarios, analizar las horas de juego y poder detectar posibles influencers del juego en base la cantidad de usuarios que encontraron útil o no su reseña.

Con esta información es posible responder las siguientes preguntas, entre otras que puedan escapar de la visión inicial de los autores: ¿Qué es lo que comúnmente más comentan los usuarios?, ¿Cómo o de qué manera se escriben las reseñas?, ¿Cuándo se escribieron?, ¿Quién o quiénes son los usuarios que podrían influenciar positiva o negativamente la percepción del juego? ¿Cuáles son los comentarios positivos o negativos más comunes? ¿Qué tendencias existen en la recomendación o no del juego en función de los atributos del conjunto de datos?.

## 8. LICENCIA.

Luego de consultar las diferentes posibilidades, la licencia utilizada fue **Creative Commons Attribution 4.0 International**, para que este trabajo pueda ser utilizado con libre acceso, pueda ser compartido, transformado para cualquier propósito incluso comercialmente sin embargo debe atribuirse dando crédito de manera adecuada a los creadores lo cual no quiere decir que el uso que se le vaya a dar tenga el apoyo de los mismos.

## 9. CÓDIGO.

Se codificó utilizando el lenguaje de programación Python y la librería Scrapy, utilizando los paradigmas de Programación orientada a objetos y procedimental, manejando de manera habitual el uso de variables y comentarios en inglés. Tomando

---

<sup>1</sup> <https://www.inputmag.com/gaming/the-fall-guys-phenomenon-explained/amp>

<sup>2</sup> <https://medium.com/the-innovation/the-psychology-of-the-fall-guys-phenomenon-fo82f594f92a>



como referencia el caso de prueba, la recolección duró aproximadamente **2 horas y 35 minutos** y se extrajeron un total de **105.247** registros.

El código fuente del proyecto se encuentra disponible en:

<https://github.com/cahucadi/GamesScraping.git>


#### 10. DATASET.

Se ha publicado el dataset en formato CSV en Zenodo y el DOI obtenido se detalla a continuación:

**10.5281/zenodo.4244834**

La URL para acceder al dataset publicado es la siguiente

**<http://doi.org/10.5281/zenodo.4244834>**

Contribuciones	Firma	Firma
Investigación previa	CHCD	
Redacción de las respuestas	CHCD	
Desarrollo código	CHCD	