

Práctica 1: Web Scraping

David Fonfría Solabarrieta

Javier Martínez Delgado

6 de noviembre de 2021

Descripción

En esta práctica se elabora un caso práctico orientado a aprender a identificar los datos relevantes para un proyecto analítico y usar herramientas de extracción de datos. El objetivo de esta actividad será la creación de un dataset a partir de los datos contenidos en un sitio web

1. Contexto

Valorant es el videojuego competitivo de tipo fps (first person shooter) de moda a día de hoy. Al ser competitivo, se realizan varios torneos alrededor del mundo en función del porcentaje de jugadores que jueguen en cada región. Cada región tiene uno o más torneos principales que se denominan “masters” por los que, dependiendo de los puntos que consiga un equipo, éste podrá optar a una plaza para el mundial que se celebrará en diciembre. Además de las masters, hay unos torneos que actuarán como última oportunidad para acceder al mundial. En estos torneos participarán equipos de regiones más amplias. Todos los datos competitivos de Valorant son recogidos en la página web de vlr.gg, que a día de hoy no tiene fichero robots.txt, no tiene API ni tampoco ofrece funcionalidades a los usuarios para descargar sus datasets, por lo que nos ha parecido una oportunidad perfecta para utilizar técnicas de web scraping

2. Título

«Datos de las principales competiciones del primer año de Valorant»

3. Descripción del dataset

Este dataset contiene los datos de los principales torneos de Valorant de todas las regiones. Se almacenan los nombres de los jugadores, los personajes (*Agentes*), diversas estadísticas del juego, así como la imagen asociada al agente.

4. Representación gráfica

masterberlin																											
Player	Country	Team	FirstAgent	SecondAgent	MoreThan2Agents	OTP	Rounds	AverageCombatScore	KillsDeaths	AverageDamagePerRound	KillsPerRound	AssistsPerRound	FirstKillsPerRound	FirstDeathsPerRound	HeadshotPercentage	ClutchesPercentage	MaxKillsPerMap	Kills	Deaths	Assists	FirstKills	FirstDeaths	Tournament	Region	ImageFirstAgent		
JohnQan	Thailand	NGLB	noxa	myra		1	0	73	304.0	1.26	197.6	1.0	0.19	0.14	0.14		25	20	29	73	58	14	10	10	Master Berlin	Europe	noxa.png
PuckTap	Thailand	X10	jett	sage		0	0	191	255.0	1.59	141.2	0.92	0.16	0.22	0.08		19	27	25	176	111	30	42	15	Master Berlin	Europe	jett.png
disponer	Philippines	ESP	noxa		0	0	1	54	254.7	1.08	162.2	0.8	0.33	0.15	0.13		0	18	43	40	18	8	7	Master Berlin	Europe	noxa.png	
ZedBauer	Singapore	SMG	jett		0	0	1	247	245.4	1.29	155.9	0.89	0.17	0.15	0.12		22	25	35	221	171	41	26	30	Master Berlin	Europe	jett.png
fozz	Thailand	X10	noxa	noxa		0	0	191	244.9	1.26	161.3	0.84	0.21	0.11	0.13		21	21	25	161	128	59	21	25	Master Berlin	Europe	noxa.png
berLIN	Taiwan	KPC	noxa		0	0	1	81	242.8	1.05	198.3	0.8	0.22	0.12	0.15		21	23	21	65	62	18	10	12	Master Berlin	Europe	noxa.png
DuBeeP	Philippines	ESP	jett		0	0	1	54	242.7	1.15	142.7	0.83	0.15	0.15	0.07		23	0	22	45	26	8	8	4	Master Berlin	Europe	jett.png
LEWY	Singapore	SMG	breach	noxa		1	0	247	237.9	1.06	142.7	0.77	0.26	0.13	0.13		15	18	23	189	179	64	23	33	Master Berlin	Europe	breach.png
mondekill	Philippines	AA	killy		0	0	1	43	233.0	0.97	139.8	0.74	0.12	0.14	0.0		19	14	20	32	33	5	6	0	Master Berlin	Europe	killy.png
budimeistaf	Indonesia	NGLB	orion	brimstone		0	0	45	215.0	0.97	148.7	0.73	0.24	0.09	0.11		28	0	19	33	34	11	4	5	Master Berlin	Europe	orion.png
Egnot	Singapore	SMG	killy	cypher		1	0	247	212.5	1.08	134.1	0.74	0.23	0.09	0.09		25	31	21	183	170	58	22	23	Master Berlin	Europe	killy.png
hickling	Indonesia	BMG	noxa	noxa		0	0	120	210.0	1.07	134.1	0.73	0.23	0.1	0.06		28	25	19	85	82	27	12	7	Master Berlin	Europe	noxa.png
iduary	Thailand	X10	viper	brimstone		1	0	191	205.8	1.11	132.3	0.71	0.29	0.11	0.07		17	24	25	136	123	56	21	14	Master Berlin	Europe	viper.png
Asterisk	Indonesia	BMG	orion	brimstone		0	0	120	203.7	1.03	137.2	0.68	0.32	0.06	0.08		17	15	17	82	80	38	7	10	Master Berlin	Europe	orion.png
Winer	Taiwan	KPC	jett		0	0	1	81	202.5	0.98	123.9	0.7	0.11	0.11	0.12		23	25	23	97	98	8	9	10	Master Berlin	Europe	jett.png
LAZERBNAZ	Thailand	NGLB	noxa		0	0	1	73	201.5	0.86	127.7	0.63	0.14	0.07	0.05		24	8	127	45	48	16	5	4	Master Berlin	Europe	noxa.png

Figura 1: Visualización del dataset obtenido

5. Contenido

El periodo de tiempo de los datos es desde el 2 de junio de 2020 hasta ahora, ya que es la fecha en la que se lanzó y dentro de un mes es el primer torneo mundial. La página de la que hacemos el scraping ha recogido los datos a mano mediante las retransmisiones de los torneos ya que por el momento, no existe ninguna API para acceder a los datos del videojuego como sí que pasa en otros videojuegos de la compañía. El dataset está compuesto de los siguientes campos:

- **Player:** Nick del jugador
- **Country:** País del jugador-Team: Equipo con el que ha participado en el torneo
- **FirstAgent:** Agente (Personaje) más utilizado por el jugador
- **SecondAgent:** Segundo agente más utilizado por el jugador
- **MoreThan2Agents:** 1 cuando el jugador ha utilizado más de dos agentes en el torneo
- **OTP:** 1 cuando el jugador es OTP (One Trick Pony, se refiere a cuando un jugador se centra en un único agente y no utiliza ninguno más)
- **Rounds:** Número de rondas jugadas
- **AverageCombatScore:** Media del Combat Score del jugador a lo largo del torneo. Se trata de una medida indicativa de lo que ha hecho el jugador en cada ronda (objetivos, muertes, asistencias, etc)
- **KillsDeaths:** Proporción muertes/bajas
- **AverageDamagePerRound:** Daño medio por ronda
- **KillsPerRound:** Muertes por ronda
- **AssistsPerRound:** Asistencias por ronda
- **FirstKillsPerRound:** Primeras muertes por ronda
- **FirstDeathsPerRound:** Primeras bajas por ronda
- **HeadshotPercentage:** Porcentaje de disparo en la cabeza

- **ClutchesPercentage**: Porcentaje de victoria en *clutches* (cuando queda un jugador contra varios del equipo enemigo)
- **MaxKillsPerMap**: Número máximo de muertes por partida
- **Kills**: Número total de muertes
- **Deaths**: Número total de bajas
- **Assists**: Número total de asistencias
- **FirstKills**: Número total de primeras muertes
- **FirstDeaths**: Número total de primeras bajas
- **Tournament**: Torneo al que pertenecen los datos
- **Region**: Región del torneo
- **ImagenFirstAgent**: Nombre de la imagen del agente más usado por el jugador

6. Agradecimientos

El propietario del conjunto de datos es la página de vlr.gg. En cuanto a los análisis anteriores, normalmente al tratarse de deportes electrónicos no se realizan análisis escritos ya que se suelen realizar en forma de vídeos o directos. Un ejemplo es un famoso analista español que sube sus análisis a la plataforma de youtube, por ejemplo [este vídeo](#)¹, que es un análisis de un partido de la máster de Berlín.

En observación de los principios éticos y legales se ha escogido un sitio web cuyos propietarios no han expresado su deseo de no ser objetivo de técnicas de web scraping así como un conjunto de datos cuyo copyright permita su descarga.

7. Inspiración

Nos ha parecido interesante este conjunto de datos para poder analizar a los equipos que participarán en el mundial y determinar qué aspectos tendría que buscar un equipo para mejorar. Como se ha dicho en el apartado anterior, nos hemos podido inspirar en los análisis que se suelen realizar partido a partido, para realizar un dataset con los datos desde que se lanzó el juego hasta la actualidad.

8. Licencia

CC0: Public Domain License. Hemos elegido esta licencia ya que no nos interesa poner ninguna restricción a su uso.

¹URL: <https://www.youtube.com/watch?v=JM7DtLAdBMU&list=PLjj8yZCK7bXP0uII2l9M6Hz7WVm2DcJ8M&index=13>

9. Código

Enlace Github: <https://github.com/ArcticAI/Practica1tipologia>

10. Dataset

10.5281/zenodo.5646234

11. Contribuciones

Contribuciones	Firma
Investigación Previa	J.M.D., D.F.S.
Redacción de las respuestas	J.M.D., D.F.S.
Desarrollo de código	J.M.D., D.F.S.