TIPOLOGÍA Y CICLO DE VIDA DE LOS DATOS PRÁCTICA 1

Luis Martin de Pablo

Universitat Oberta de Catalunya

Presentación

En esta práctica se elabora un caso práctico orientado a aprender a identificar los datos relevantes por un proyecto analítico y usar las herramientas de extracción de datos. Tendréis que entregar un solo fichero con el enlace Github donde haya las soluciones incluyendo los nombres de los componentes del equipo. Podéis utilizar la Wiki de Github para describir vuestro equipo y los diferentes archivos de vuestra entrega.

Competencias

En esta **PEC** se desarrollan las siguientes competencias del **Máster de Data Science**:

- Capacidad de analizar un problema en el nivel de abstracción adecuado a cada situación y aplicar las habilidades y conocimientos adquiridos para abordarlo y resolverlo.
- Capacidad para aplicar las técnicas específicas de web *scraping*.

Objetivos

Los objetivos concretos de esta **PEC** son:

- Aprender a aplicar los conocimientos adquiridos y su capacidad de resolución de problemas en entornos nuevos o poco conocidos dentro de contextos más amplios o multidisciplinarios.
- Saber identificar los datos relevantes que su tratamiento aporta valor a una empresa y la identificación de nuevos proyectos analíticos.
- Saber identificar los datos relevantes para llevar a cabo un proyecto analítico.
- Capturar datos de diferentes fuentes de datos (tales como redes sociales, web de datos o repositorios) y mediante diferentes mecanismos (tales como *queries*, API y *scraping*).
- Actuar con los principios éticos y legales relacionados con la manipulación de datos en función del ámbito de aplicación.
- Desarrollar la capacidad de búsqueda, gestión y uso de información y recursos en el ámbito de la ciencia de datos.

Scraper

Contexto

Todo el mundo conoce la exitosa franquicia de Pokémon. Empezó en 1996 como un videojuego RPG con los lanzamientos de *Pokemon Rojo y Verde* en Japón. Desde entonces y gracias a su creciente popularidad se ha extendido a otros medios como series de televisión, películas y juegos de cartas, convirtiéndose en una marca reconocida a nivel mundial. La empresa a cargo del desarrollo de los videojuegos es Game Freak, y la distribución corre a cargo de Nintendo.

La mecánica principal de estos videojuegos consiste en tomar el papel de un entrenador Pokémon y hacer luchar tus Pokémon con los del entrenador adversario. Tal es el éxito de esta franquicia que a día de hoy se organizan torneos a nivel mundial de entrenadores Pokémon. La naturaleza competitiva de esta mecánica hace que uno de los papeles más importantes de estos videojuegos sea la de formar el mejor equipo Pokémon posible.

El lanzamiento del primer videojuego en 1996 incluía un total de 151 Pokémon distintos. Veinticinco años más tarde existen un total de 898 Pokémon distintos, cada uno con sus estadísticas, sus habilidades, su set de movimientos y sus tipos. Todo esto sin contar las formas o variaciones presentes en algunos Pokémon. Queda vigente pues, la necesidad de recopilar toda esta información con el objetivo de estudiar mejor las diferentes estrategias posibles para convertirse en el mejor entrenador Pokémon.

Actualmente existen múltiples páginas web que se encargan de recopilar en miles de tablas todos los datos de todos los Pokémon. Los usuarios pueden acceder a ellas en caso de necesitar alguna información especifica, cómo podría ser consultar qué ataques puede aprender determinado Pokémon o cuáles son sus estadísticas. Sin embargo, sacar métricas a partir de la totalidad de la Pokédex (conjunto de todos los Pokémon) para saber qué tipo tiene en promedio el ataque más alto es algo más complicado. Con este fin se ha recopilado la información más básica de la Pokédex mediante web scraping de la página web https://pokemondb.net.

Descripción

En esta práctica se han recopilado los datos mas relevantes de la Pokédex mediante web scraping de la página web https://pokemondb.net, más concretamente de https://pokemondb.net/pokedex/all. Toda la información recogida se ha guardado en el dataset de nombre **National Pokédex**, que hace referencia a la totalidad de Pokémon existentes. El dataset reúne las características principales de los 893 Pokémon así como de todas sus formas y variaciones, lo que suman un total de 1045 entradas. Entre estos datos se encuentran el tipo y las estadísticas base de cada Pokémon. De manera paralela también se han recogido el arte y el icono de cada Pokémon y guardado en carpetas independientes al dataset.

La información más relevante de este dataset son el tipo y las estadísticas de los distintos Pokémon. El número de la Pokédex y el nombre nos permiten identificar el Pokémon en cuestión, pero la descripción de la especie, la altura y el peso son datos más irrelevantes. El peso influye en algunos ataques, en que el daño o la velocidad del Pokémon varía en función del peso, pero son ataques muy específicos útiles simplemente para los Pokémon más pesados o más ligeros. Estos datos se recogen por completitud del dataset o para realizar otro tipo de estudios no relacionados con el mundo competitivo de Pokémon.

El primer dato importante son los tipos. Cada Pokémon tiene asociada una combinación de tipos que define en gran medida los ataques que podrá usar, la efectividad de estos ataques, y sus debilidades y resistencias. Cada ataque tiene a su vez un tipo. Se asemeja al juego clásico de piedra, papel tijeras. Un Pokémon de tipo agua será débil frente a un ataque de tipo planta, un Pokémon de tipo planta será débil ante un ataque de tipo fuego, y un Pokémon de tipo fuego será débil frente un ataque de tipo agua. Un Pokémon de tipo agua aprenderá mayormente ataques de tipo agua, pero podría aprender un movimiento de otro tipo. Cuando se realiza un ataque del mismo tipo del Pokémon, el poder de este ataque se multiplica. A todo esto, el tipo de un Pokémon puede ser la combinación de dos tipos (planta - veneno, por ejemplo). Lo que genera una combinación de debilidades y resistencias distinta, y proporciona un rango más amplio de tipos de ataque.

El segundo dato importante que recoge el dataset son las estadísticas base del Pokémon. Como bien dice el nombre, las estadísticas base son las estadísticas que tiene un Pokémon al nacer o evolucionar. Un Pokémon se puede entrenar para fortalecer determinadas estadísticas, pero las estadísticas base definen cuál es el máximo potencial del Pokémon. Son el esqueleto sobre el cual se construyen la estadísticas del Pokémon. Se dividen en seis: los puntos de vida, que reflejan la constitución de un Pokémon, la cantidad de daño que puede resistir; la potencia de ataque, solo útil para los ataques de origen físico, ataques basados en fuerza muscular; la defensa, indica la resistencia frente a ataques de tipo físico; la potencia de ataque especial, útil para ataques no basados en la fuerza muscular, como podría ser ataques de origen psíquico o ataques de origen elemental, como el fuego y el agua; la defensa especial, indica la resistencia frente ataques de tipo especial; la velocidad, en un combate, el Pokémon más veloz ataca primero. Cabe destacar que dos Pokémon idénticos, con las mismas estadísticas base, entrenados al máximo, no tendrán necesariamente las mismas estadísticas, ya que podemos escoger qué aspectos del Pokémon queremos potenciar. El entrenamiento del Pokémon también es un aspecto fundamental. Generalmente se entrenan los puntos fuertes del Pokémon, es decir, las estadísticas base más altas.

Antes de efectuar el scraping se ha comprobado que el servidor no tuviese a la disposición de los usuarios ninguna API que facilite el proceso de recogida de datos, y se ha comprobado el archivo robots.txt. Se ha intentado, sin éxito, cumplir automáticamente las guías de comportamiento establecidas por la web de manera automática mediante un robots.txt parser. Es posible que la codificación del fichero de esta página web no siga con rigor las reglas establecidas para este tipo de archivos. De todos modos se ha verificado el cumplimiento de estas guías de manera manual. El esbozo del código del robots.txt parser a quedado comentado en el código.

El web scraper se sirve de un crawler que se encarga de efectuar la petición http y descargar el contenido de la página mediante la librería requests. Utilizando la librería BeautifulSoup se recorre la url principal para acceder a la tabla que contiene la gran mayoría de los datos y recorriendo esta tabla se accede a los links propios de cada entrada para recoger información más específica.

Diagrama del scraping y CSV

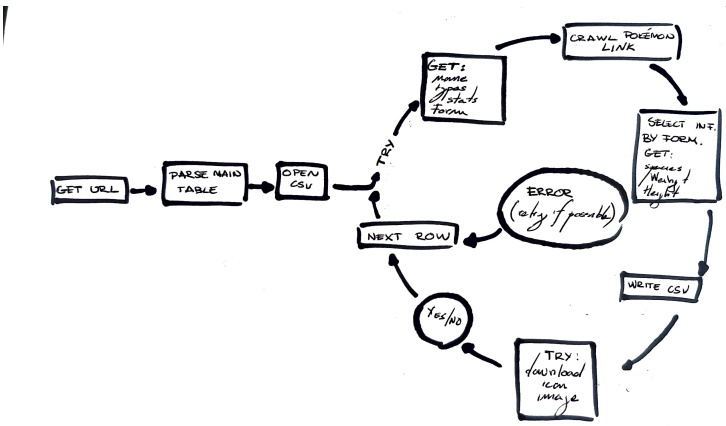
El siguiente esquema muestra un resumen del proceso de obtención de datos. El primer paso es descargar la página web https://pokemondb.net/pokedex/all utilizando el crawler, una función definida a parte que permite descargar la web mediante una petición http utilizando la librería requests. Esta función pide la url como parámetro de entrada, y permite la modificación del user agent. En caso de error la función vuelve a reintentar el proceso, siempre y cuando se deba a un error de conexión o un error del servidor (5xx), dejando transcurrir un intervalo de tiempo para permitir la recuperación de la conexión o del dominio web. Por defecto el número de reintentos es dos, pero este parámetro también se puede modificar. También se podría añadir un robot parser como parámetro de entrada, que descargase el dominio siguiendo las guías del archivo robots.txt. En nuestro caso no ha sido posible debido a que el archivo no sigue las reglas de escritura estándar.

Tras descargar la web principal, utilizando la librería BeautifulSoup se guarda la tabla con la información relevante y se navega por ella, fila a fila mediante un bucle, para recopilar los datos de interés. El proceso de extracción de datos de cada fila se encuentra dentro de un try, para de este modo saltar al siguiente registro en caso de encontrarse con algún error inesperado. De cada registro de la tabla principal se guarda el número de la Pokédex, el nombre del Pokémon, sus estadísticas, y se comprueba si s trata de una variación o forma alternativa de otro Pokémon. Aunque más adelante, de esta página también se descarga el icono .png.

Posteriormente se descarga, de una manera similar, la web específica asociada al Pokémon de ese registro. Dado que las diferentes variaciones de un Pokémon comparten página web, se tiene que seleccionar a qué información queremos acceder. Se puede acceder a la información utilizando el código que la web asocia a cada Pokémon o forma. De esta segunda tabla se extrae la descripción de la especie, la altura y el peso. Se escribe toda la información en el CSV y se procede a descargar las imágenes. Esta descarga, tanto del icono de la tabla principal como de la imagen en la web del Pokémon, se efectúa dentro de otro try para poder proseguir con el scraping en caso de que la descarga de las imágenes falle. Las imágenes se guardan en sus carpetas correspondientes, creadas al inicio del programa.

Finalmente se reinicia el bucle saltando al siguiente registro. Se espera un tiempo prudencial entre diferentes iteraciones para evitar saturar al servidor web de peticiones http. Una vez recorrida toda la tabla se cierra el archivo CSV.

ID	Number	Name	Type1	Type2	Total	HP	Atk	Def	SpAtk	SpDef	Spd	Species	Height	Weight
1	1	Bulbasaur	Grass	Poison	318	45	49	49	65	65	45	Seed Pokémon	0.7 m	6.9 kg
2	2	Ivysaur	Grass	Poison	405	60	62	63	80	80	60	Seed Pokémon	1.0 m	13.0 kg
3	3	Venusaur	Grass	Poison	525	80	82	83	100	100	80	Seed Pokémon	2.0 m	100.0 kg
4	3	Venusaur(Mega Venusaur)	Grass	Poison	625	80	100	123	122	120	80	Seed Pokémon	2.4 m	155.5 kg



Ejemplo de uso

Se quiere construir un equipo alrededor de nuestro Pokémon favorito, Charizard. Charizard será nuestro Pokémon principal y se quiere averiguar cuál podría ser su compañero ideal en un combate doble. Al ser un Pokémon de tipo fuego - volador, sus debilidades principales son los tipos agua, eléctrico y roca. Queremos un Pokémon que supla estas debilidades. El tipo Planta es eficaz contra agua y roca, y resistente al tipo eléctrico. Por otro lado, Charizard es un atacante especial. Queremos un Pokémon atacante físico para complementarlo mejor. Usando nuestra base de datos se puede encontrar al Pokémon tipo planta con la estadística de ataque físico más alta, Rillaboom.

Este ejemplo es una muestra muy simple de las consultas que podemos hacer. Se podrían realizar consultas algo más elaboradas, como por ejemplo, buscar Pokémon que tengan la suma de sus estadísticas defensivas lo más alta posible. El uso de la base de datos es muy sencillo:

1. Escoger las características deseadas.

2. Buscar los Pokémon que cumplen estas características.

Contenido

Para cada uno de los Pokémon se recoge la siguiente información:

- **Number:** Número de la Pokédex. Identificativo de cada Pokémon. No se trata de una clave primaria dado que las formas alternativas de los Pokémon, aún teniendo estadísticas distintas, comparten número con la forma base.
- **Name:** Nombre del Pokémon.
- **Type1:** Tipo principal del Pokémon. (Planta, Agua, Fuego, etc.)
- **Type2:** Tipo secundario del Pokémon, en caso de tener. (Planta, Agua, Fuego, etc.)
- **Total:** Suma de todas las estadísticas base del Pokémon. (HP, Atk, Def, SpAtk, SpDef, Spd)
- **HP:** Estadística base de **vida** del Pokémon.
- **Atk:** Estadística base de **ataque** del Pokémon.
- **Def:** Estadística base de **defensa** del Pokémon.
- **SpAtk:** Estadística base de **ataque especial** del Pokémon.
- **SpDef:** Estadística base de **defensa especial** del Pokémon.
- **Spd:** Estadística base de **velocidad** del Pokémon.
- **Species:** Especie del Pokémon.
- **Height:** Altura en metros del Pokémon.
- **Weight:** Peso en kilos del Pokémon.

La fuente de toda esta información es originalmente el conjunto de videojuegos de Pokémon. La página web en cuestión lleva proporcionando estos datos y otro contenido relacionado con Pokémon desde 2008, actualizándose a medida que salen al mercado nuevos videojuegos. La web utiliza el proyecto Veekun para obtener parte de los datos, un proyecto sin ánimo de lucro que se encarga de hacer scraping sobre el código base de los videojuegos y se apoya en la comunidad para corregir posibles errores.

Cada tres o cuatro años el universo de Pokémon suele expandirse, añadiendo nuevos Pokémon a la Pokédex. Cuando eso suceda este dataset pasará a estar incompleto. Normalmente los Pokémon antiguos no sufren ningún cambio en sus estadísticas, aún así no se puede garantizar que todos los registros seguirán siendo vigentes.

Agradecimientos

La totalidad de los datos del dataset han sido recolectados de la página web https://pokemondb.net. Parte de la información de la web proviene del proyecto Veekun. Ambos proyectos creados y mantenidos por una única persona con la ayuda de la comunidad de Pokémon. En última instancia, todos los derechos sobre la información recopilada pertenece a ©1995-2021 Nintendo/Game Freak.

Como se ha mencionado en la introducción, existen una gran cantidad de páginas web, creadas y mantenidas por la comunidad de fans de la saga, que se encargan de recopilar toda la información relacionada con el mundo de Pokémon, información generalmente usada con motivos de ocio. WikiDex, Bulbapedia o serebii son algunos ejemplos. Estas páginas web están enfocadas a la búsqueda de información específica, no están pensadas para la extracción de métricas y estadísticas. Existen otras páginas mucho más enfocadas al mundo competitivo de Pokémon, como podría ser smogon. Esta web se encarga de estudiar las mejores estrategias de cada Pokémon, dividir los Pokémon en rangos según su porcentaje de victorias y versatilidad en combate, así como proporcionar una gran cantidad de información relacionada con el competitivo. Aun así, no nos proporciona las herramientas para poder sacar nuestras propias estadísticas.

El objetivo de este dataset se asemeja más proyectos como el mencionado anteriormente proyecto Veekun. Este proyecto se encarga de hacer scraping sobre el código de los videojuegos, para extraer toda la información posible de ellos. Toda esta información está recopilada en un repositorio y es una de las principales fuentes para todas las páginas web mencionadas previamente.

Inspiración

La marea creciente del fenómeno de Pokémon es algo indiscutible. La naturaleza competitiva de este videojuego hace que un gran porcentaje de jugadores recurra a páginas web especializadas, con el objetivo de elaborar estrategias y crear el mejor equipo Pokémon posible. El objetivo de este dataset es recopilar la información más básica de todos los Pokémon con el fin de permitir consultas más elaboradas que no se podrían hacer en la interfaz de una página web. Se trata de un dataset enfocado principalmente al ocio, pero que en caso de necesidad podría usarse como herramienta para ganar el premio gordo de los multitudinarios torneos de Pokémon que se realizan alrededor del mundo.

Esta primera versión del dataset recoge las principales características necesarias para la elaboración de estrategias en el mundo de Pokémon. En versiones futuras sería muy interesante añadir una tabla a cada Pokémon que recopilase todos los ataques que puede aprender, con su descripción, su efecto, el tipo y el modo de aprenderlo. El conjunto de todas estas tablas proporcionaría todas las herramientas necesarias para desarrollar estrategias elaboradas sin la necesidad de recurrir a páginas especializadas como smogon, que te ofrecen la información, pero no la capacidad de realizar tus propios estudios.

Licencia

La política de distribución de este dataset se ampara bajo la Licencia CC BY-NC-SA 4.0, que permite compartir y adaptar la información bajo las siguientes restricciones:

- **Reconocimiento:** Debe reconocer adecuadamente la autoría, proporcionar un enlace a la licencia e indicar si se han realizado cambios. Puede hacerlo de cualquier manera razonable, pero no de una manera que sugiera que tiene el apoyo del licenciador o lo recibe por el uso que hace.
- **No Comercial:** No puede utilizar el material para una finalidad comercial.
- **Compartir Igual** — Si remezcla, transforma o crea a partir del material, deberá difundir sus contribuciones bajo la misma licencia que el original.

Dado que los nombres y las imágenes de todos los Pokémon así como la información asociada son propiedad de ©1995-2021 Nintendo/Game Freak, se prohibe el uso de este dataset para fines comerciales. No obstante, se puede modificar y distribuir libremente indicando los cambios realizados y otorgando reconocimiento al creador del dataset original.

Publicación

El fichero CSV ha sido publicado en la web Zenodo con el DOI **10.5281/zenodo.4665380**. Se puede acceder a el mediante la siguiente url: https://zenodo.org/record/4665380.

El proyecto está público en el siguiente repositorio de GitHub: https://github.com/luismartindepablo/Pokemon_WebScraping.git.

Bibliografía

[1] Subirats, L., Calvo, M. (2019). Web Scraping. Editorial UOC.

[2] Lawson, R. (2015). Web Scraping with Python. Packt Publishing Ltd. Chapter 2. Scraping the Data.

[3] Pokémon Database (2021) Retrieved 6 April 2021, from https://pokemondb.net

[4] veekun/pokedex (2021) Retrieved 6 April 2021, from https://github.com/veekun/pokedex

[5] WikiDex (2021) Retrieved 6 April 2021, from https://www.wikidex.net/wiki/WikiDex

[6] Bulbapedia, the community-driven Pokémon encyclopedia (2021) Retrieved 6 April 2021, from https://bulbapedia.bulbagarden.net/wiki/Main_Page

[7] Serebii.net - Where Legends Come To Life (2021) Retrieved 6 April 2021, from https://www.serebii.net

[8] Smogon University - Competitive Pokémon Community (2021) Retrieved 6 April 2021, from https://www.smogon.com