Tipología y ciclo de vida de los datos

PRÁCTICA 1: Web scrapping

Autores: Eleazar Morales Díaz y Susana Vila Melero

Marzo 2021

Contents

1. Contexto.

El conjunto de datos recoge información sobre diferentes ejercicios de programación creados con la intención de fomentar o desarrollar habilidades y competencias. A dichos ejercicios se les denomina "katas".

Utilidades

- Obtener información sobre el interés que suscitan los diferentes lenguajes de programación en el mercado.
- Análisis estadístico descriptivo del sector del desarrollo del software.
- Comparativa entre rendimiento, potencia y evolución de los diferentes lenguajes de programación.
- Desde el punto de vista empresarial, acceso a información valiosa sobre perfiles expertos en los lenguajes de interés.

2. Definir un título para el dataset.

Katas (ejercicios de programación) y estadísticas asociadas.

3. Descripción del dataset.

Conjunto de datos que recolecta información sobre las katas de programación recientemente resueltas en la plataforma `CodeWars`, con diferentes lenguajes y niveles de dificultad.

4. Representación gráfica.

Figure 1: Primera posibilidad

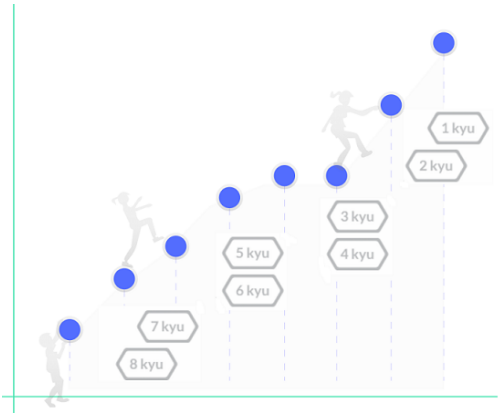

Figure 2: Segunda posibilidad

5. Contenido.

El dataset recoge información de las katas resueltas recientemente, aplicando el filtro *Approved*. El número total de katas en esa categoría asciende a 6.932 en la fecha de raspado el 18 de marzo de 2021.
No se dispone de información sobre la fecha de ejecución de la kata, pero sí se conoce la fecha de publicación de esta. Teniendo en cuenta ese dato, el dataset contiene katas publicadas entre el mes de marzo de 2013 y el mes de marzo de 2021.

Para cada una de estas katas, se incluye una línea con los campos que se listan a continuación.

- **id_kata**: identificador único de la kata.
- **name**: nombre de la kata.

- **author**: autor de la kata.
- **author_profiles**: perfiles sociales (si existen) del autor de la kata.
- **tags**: tags de la kata.
- **kata_complexity**: complejidad de resolución de la kata; es un valor que va desde *8 kyu* (la más sencilla) a *4 dan* (la más compleja).
- **published**: fecha de publicación de la kata en la página web.
- **warriors_trained**: número de personas que han entrenado con la kata.
- **total_skips**: número de personas que han abandonado la kata sin resolverla.
- **total_code_submissions**: total de soluciones presentadas.
- **total_times_completed**: número de veces que la kata ha sido resuelta.
- **languages_completions**: número de veces que la kata ha sido resuelta por lenguaje de programación.
- **total_stars**: número de veces que se ha agregado la kata a favoritos.
- **positive_feedback**: Porcentaje de votos positivos sobre el total.
- **total_very_satisfied_votes**: Número de votantes muy positivos.
- **total_somewhat_satisfied_votes**: Número de votantes positivos.
- **total_not_satisfied_votes**: Número de votantes insatisfechos.
- **total_rank_assessments**: número de rangos que han participado en la kata.
- **average_assessed_rank**: rango medio de los que han entrenado con la kata.
- **highest_assessed_rank**: rango más elevado de los que han entrenado con la kata.
- **lowest_assessed_rank**: rango inferior de los que han entrenado con la kata.

Se debe tener en cuenta que ciertas columnas forman una lista variable de elementos, como puede pasar con las columnas **author_profiles**, **tags** y **languages_completions**. En estos casos se usa la siguiente notación:

```
author_profiles = [author_1_uri, author_2_uri]

tags = [tag_1, tag_2]

languages_completions = [(language_1, total_times_completed), (language_2, total_times_completed)]
```

6. Agradecimientos.

Los datos han sido obtenidos de la plataforma Codewars. Se trata de una comunidad web que nace como el esfuerzo colaborativo de usuarios que desinteresadamente aportan katas de entrenamiento, soluciones a las mismas y feedback constructivo. Este servicio ayuda a crecer profesionalmente a la comunidad de desarrolladores en su campo de conocimiento

Hemos encontrado otros análisis similares relativos a lenguajes de programación, por ejemplo el que publica TIOBE Software mediante su **indicador de popularidad de lenguajes de programación** con periodicidad mensual, o la **Developer Survey** que publica anualmente Stack Overflow

7. Inspiración.

Los estudios citados en el apartado anterior permiten conocer la evolucion temporal del uso de los lenguajes de programación. El dataset obtenido proporciona además respuesta a las siguientes preguntas:

- ¿Qué lenguajes de programación son los preferidos para competir?
- ¿Cual es el reparto porcentual de usuarios por lenguaje de programación combinando más de uno de ellos?
- ¿Cómo se distribuyen los desarrolladores según la complejidad de las katas?
- ¿Qué katas resultan más atractivas y porqué?

- ¿Soy una empresa de desarrollo de software o un profesional en RRHH ¿Qué perfiles son los más interesantes para la contratación?

8. Licencia.

Se ha decidido hacer uso de la licencia **GPLv3** recomendado por la Free Software Foundation. (FSF). Nos interesa especialmente por las libertades que la misma ofrece a los usuarios. Se permite su distribución, se reconoce el autor de la obra, se permite editar el código fuente, incluso lucrarse economicamente con el mismo. No obstante, no se permite privatizar el software con una licencia que altere las libertades anteriormente expuestas.

9. Código.

El código utilizado para la extracción de la información requerida mediante Web Scrapping se puede acceder mediante la siguiente URL.

10. Dataset.

El dataset se ha publicado en formato CSV en Zenodo DOI con la siguiente descripción:

The dataset contains information extracted from the website "Codewars" about different programming exercises created with the intention of fostering or developing abilities and competences. Said exercises are called "katas". Our dataset collects information from recently solved katas and their statistics, applying the "Newest" filter. The total number of katas in that category amounts to 6.931. Information about the date of execution is not available beyond the previously mentioned filter, but the date of publication of the kata is known. Taking this factor into account, the dataset contains katas published between March 2013 and March 2021.

Además se puede leer en la carpeta **data** bajo el siguiente enlace.

Tabla contribuciones

Contribuciones	**Firma**
Investigación Previa	Eleazar Morales Díaz, Susana Vila Melero
Redacción de las respuestas	Eleazar Morales Díaz, Susana Vila Melero
Desarrollo del código	Eleazar Morales Díaz, Susana Vila Melero