Universitat Oberta de Catalunya
Master de Ciencia de Datos
Diseño y Uso de Bases de Datos Analíticas

# PRA 1

Integrante: Eduardo Béjar Feijoó
Fecha: 8 de noviembre de 2021

## 1. Contexto.

Para esta práctica de Web Scraping estuve analizando varios temas y sitios web para utilizar como fuente; finalmente seleccioné el tema de montañismo por una afición personal y el sitio web Summit Post (https://summitpost.org/), sitio creado de forma colaborativa por aficionados al montañismo en el cual publican información y datos sobre montañas, rutas y actividades de exteriores.

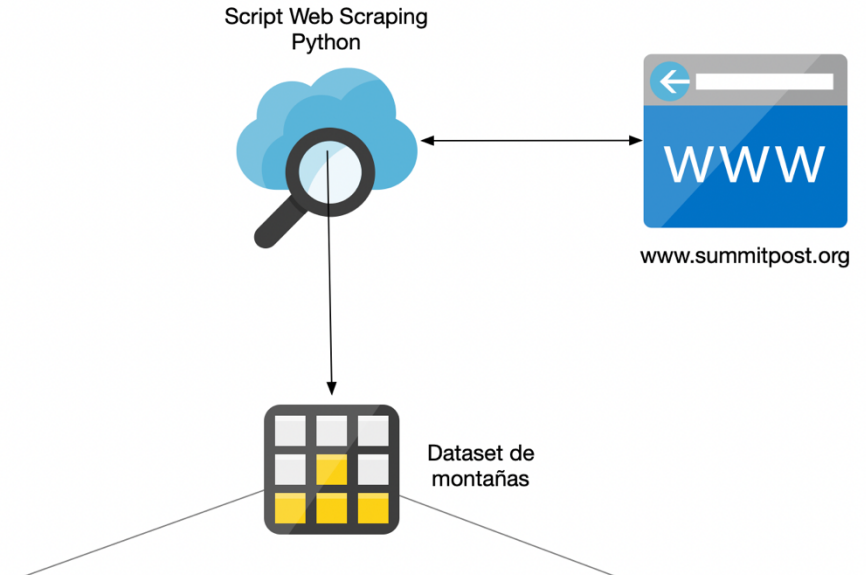
## 2. Título.

Base de montañas y elevaciones publicada en el sitio web Summit Post (https://summitpost.org/).

## 3. Descripción del dataset.

Lista de montañas y elevaciones publicadas en Summit Post (https://summitpost.org/) hasta el 8 de noviembre de 2021.

## 4. Representación gráfica.




| No. | Nombre_montana | Altura | Score | Vistas | Ubicacion |
|---|---|---|---|---|---|
| 1 | Olympus Mons | 27000 | 99.37% | 45534 | MARS,Other |
| 2 | Everest | 8849 | 100% | 253765 | Tibet/Khumbu,China/Nepal,Asia |
| 3 | K2 | 8611 | 99.85% | 260834 | China/Pakistan,Asia |
| 4 | Kangchenjunga | 8586 | 99.92% | 141388 | India/Nepal,Asia |
| 5 | Lhotse | 8516 | 95.41% | 81562 | Asia |
| 6 | Makalu | 8463 | 96.38% | 111107 | Asia |
| 7 | Cho Oyu | 8188 | 99.62% | 156312 | China/Nepal,Asia |
| 8 | Dhaulagiri | 8167 | 95.46% | 70975 | Pokhara,Nepal,Asia |
| 9 | Manaslu | 8156 | 94.89% | 68995 | GurkhaHimal,Nepal,Asia |
| 10 | Nanga Parbat | 8126 | 99.44% | 182010 | TheNorthernAreas,Pakistan,Asia |
| 11 | Annapurna I | 8091 | 94.77% | 123522 | Nepal,Asia |
| 12 | Gasherbrum I | 8068 | 98.46% | 75622 | Xinjiang/Baltistan,China/Pakistan,Asia |
| 13 | Broad Peak | 8047 | 99.62% | 103190 | Xinjiang/Baltistan,China,Pakistan,Asia |
| 14 | Gasherbrum II | 8032 | 97.41% | 68516 | Pakistan,Asia |
| 15 | Shisha Pangma | 8027 | 96.82% | 88230 | Tibet,China,Asia |

## 5. Contenido.

El conjunto de datos incluye los siguientes campos:

- **No.:** Número secuencial.
- **Nombre:** Nombre de la montaña o elevación.
- **Altura:** Altura en metros, cuando esté disponible.
- **Score:** Puntaje calculado por el sitio Summit Post en base a una fórmula que incluye la cantidad y calida de votos recibidos por los visitantes del sitio, ponderados en base al perfil de los usuarios.
- **Vistas:** Cantidad de visitas recibidas en la página de la elevación.
- **Ubicación:** Ubicación de la montaña o elevación, en términos de continentes o países.

## 6. Agradecimientos.

El conjunto de datos fue generado por Eduardo Béjar Feijoó a partir de datos publicados en el sitio web Summit Post (https://summitpost.org/), el cual es actualizado de forma colaborativa por sus usuarios. El sitio web no dispone de API para consultar los datos, tampoco se encontró un dataset similar, haciendo consultas en Google, y revisando los repositorios Zenodo y Kaggle, por ello se creó el código en Python disponible en el repositorio Github indicado.

Para el desarrollo del código primero se hizo una revisión manual del sitio web Summit Post, para identificar su estructura y código para encontrar etiquetas que puedan ser útiles para una revisión automática del sitio.

Luego se revisaron los términos y condiciones del sitio (https://www.summitpost.org/tos), y se encontró que, aunque se hace mención a que ningún material o contenido del sitio puede ser reproducido por otros sitios, el contenido es propiedad de sus autores y no de Summit Post, por lo que eventualmente se debería solicitar permiso a cada uno de los usuarios que publicaron el contenido. En el caso el código de Web Scraping, este solo extrae datos generales de las montañas (nombre, altura, ubicación), que son de conocimiento público y que están agregados en el sitio, por lo que no son creaciones intelectuales de quienes los publicaron.

## 7. Inspiración.

El conjunto de datos me resultó interesante por afición personal al montañismo. Considero que estos datos pueden ser de utilidad para definir circuitos de escalada, por ejemplo, similar al conocido circuito de los "ochomiles" (https://es.wikipedia.org/wiki/Ochomil) a partir de estos datos se podrían identificar cuáles montañas corresponden a circuitos similares como los "cuatromiles", "cincomiles" y "seismiles", para definir en un país cuál debe ser la lista de montañas a las que debe ascener alguien que quiere llegar a la cumbre de todas las que tienen una altura desde los 5.000 hasta los 5.999 metros ("cincomiles"). De esta manera podría establecer metas personales para ascender a todas las que cumplan el parámetro.

## 8. Licencia.

Para la publicación del dataset elegí la licencia Creative Commons CC-BY 4.0, buscando que las personas que los utilicen citen al fuente.

## 9. Código.

https://github.com/edobejar/UOCTIpologiaPRA1

## 10. Dataset

https://zenodo.org/record/5655408
https://doi.org/10.5281/zenodo.5655408

| Contribuciones | Firma |
|---|---|
| Investigación previa | EB |
| Redacción de las respuestas | EB |
| Desarrollo del código | EB |

**Nota sobre trabajo en grupo:**

En relación al grupo de trabajo para la práctica, inicialmente recibí confirmación de otro estudiante para conformar el grupo; sin embargo, luego recibí su excusa por lo que avancé con el trabajo de forma individual.