Pontificia Universidad Católica de Chile
Facultad de Ciencias Económicas y Administrativas

Análisis de Big Data:
EAA361A

Laboratorio 2

Profesor: Cristián Vásquez Ayudante: Pablo González

Ejercicio 1: Repaso flash (para calentar motores)

Luego de una larga travesía, Ash Ketchum (**サトシ** para los amigos) vuelve a casa con su Pokédex completa. Ésta es luego entregada al profesor Oak para su posterior análisis, quien le pide a usted, que, utilizando sus bastos conocimientos de SQL, responda las siguientes preguntas:

1. Obtenga el número, nombre, nombre en japonés, clasificación y si es o no legendario, de aquellos Pokémon cuyo ataque triplique su defensa.
2. ¿Cuáles son los nombres (en ambos idiomas) de los 10 Pokémon con una mayor felicidad base? ¿Son estos legendarios? (Responda "Sí" o "No")
3. De aquellos Pokémon legendarios, ¿qué porcentaje pertenece a cada generación?

** Datos extraídos de www.kaggle.com/rounakbanik/pokemon/data con fecha 23/03/2020*

Ejercicio 2: Datos de inmigración

A continuación, encontrará una base de datos con información acerca de la inmigración en Chile entre los años 2005 y 2016. En relación con ésta, responda las siguientes preguntas:

1. ¿Cuántas visas fueron solicitadas durante este período?
2. Calcule la edad promedio de los postulantes, al momento de hacer la solicitud.
3. Calcule el promedio de edad (actual) de los solicitantes de visa, separándolos según el año de solicitud. Recuerde que la edad es un número entero.
4. Muestre en una tabla el total de visas entregadas entre los años 2005 y 2016, separándolas según su tipo.
5. Calcular la cantidad de solicitudes y edad promedio por nivel de estudios

** Datos extraídos de https://en.datachile.io/about/data con fecha 23/03/2020*

Ejercicio 3: Recomendación de películas

Utilizando la versión ligera de "MovieLens", una base de datos con información de calificaciones y etiquetados para aproximadamente 10.000 películas, responda las siguientes preguntas:

1. Genere una nueva tabla, separando el título y año de la película en distintas variables.
2. Utilizando una medida arbitraria $\mu_{Rating} \times Log(N° \, Ratings)$, muestre los títulos y puntajes obtenidos por película, ordenándolos descendentemente.
3. Obtenga los nombres de las 10 películas con mayor cantidad de votaciones en la categoría comedia.
4. Obtenga los títulos, calificación promedio y géneros de las 10 mejores películas de animación no infantil producidas dentro de los últimos 3 años. Utilice solo el rating.
5. Muestre el título, etiqueta más común y rating promedio de aquellas 10 películas con un etiquetado más consistente (con más repeticiones).
6. **Propuesto 1:** Obtenga las 50 películas con menor calificación promedio según los 3 usuario que han escrito más reseñas.
7. **Propuesto 2:** Utilizando la medida arbitraria descrita en el ejercicio 1 y, considerando las 100 mejores películas según ésta, obtenga las 5 etiquetas más repetidas en sus descripciones.

** Datos extraídos https://grouplens.org/datasets/movielens/ con fecha 28/03/2020*