UNIVERSIDAD AUTÓNOMA DE NUEVO LEÓN

FACULTAD DE CIENCIAS FÍSICO MATEMÁTICAS

MINERÍA DE DATOS

ANÁLISIS DE LAS BASES DE DATOS

MIRTHALA NALLELY CANTÚ CORTINA

1614768

14-OCTUBRE-2020

Nombre de la base de Datos:

Google Play Store Apps

La página de iTunes App Store implementa una estructura similar a un apéndice bien indexada para permitir un raspado web simple y fácil. Por otro lado, Google Play Store utiliza sofisticadas técnicas modernas (como la carga dinámica de páginas) utilizando JQuery, lo que hace que el scraping sea más desafiante.

Cada aplicación (fila) tiene valores de categoría, clasificación, tamaño y más.

Esta información se extrae de Google Play Store. La información de esta aplicación no estaría disponible sin ella.

Objetivo: Los datos de las aplicaciones de Play Store tienen un enorme potencial para impulsar al éxito a las empresas de creación de aplicaciones. Se pueden extraer conocimientos prácticos para que los desarrolladores trabajen y capturen el mercado de Android.

Problema Planteado: En esta base de datos se habla de un público inmensamente grande y demasiado variado, por lo que resulta difícil el crear nuevo contenido y repartirlo de acuerdo a cada tipo de público.

Solución: Crear una herramienta que nos permita clasificar toda la información disponible, así como las apps que ya están a disposición del público, ver cuáles han tenido éxito y cuáles solo se han estado des-instalando, innovar en cuestión de nuevo contenido (nuevas Apps) que nos puedan aumentar aún más el público.

Nombre de la base de Datos:

Novel Corona Virus 2019 Dataset

La información de nivel diario sobre las personas afectadas puede brindar información interesante cuando se pone a disposición de la comunidad científica de datos en general.

La Universidad Johns Hopkins ha creado un excelente tablero con los datos de los casos afectados. Los datos se extraen de las hojas de Google asociadas.

Contenido

El nuevo coronavirus 2019 (2019-nCoV) es un virus (más específicamente, un coronavirus) identificado como la causa de un brote de enfermedad respiratoria detectado por primera vez en Wuhan, China. Al principio, muchos de los pacientes en el brote en Wuhan, China, según se informa, tenían algún vínculo con un gran mercado de mariscos y animales, lo que sugiere una propagación de animal a persona. Sin embargo, se informa que un número creciente de pacientes no ha estado expuesto a los mercados de animales, lo que indica que se está produciendo una propagación de persona a persona.

Objetivo: Conocer más sobre la enfermedad, de esta manera, se pueden ir recopilando todos los datos que vayan surgiendo, hacer comparaciones y clasificaciones sobre el tipo de infección, síntomas, edad de las personas, etc. Para de esta forma poder llegar a encontrar una vacuna o las precauciones máximas necesarias para salvaguardar la vida de las personas.

Planteamiento del Problema: El problema principal es que es un virus nuevo y se propaga de manera rápida entre los individuos, al ser nuevo, no tenemos forma de atacarlo y solo queda estudiarlo un poco más para poder vencerlo.

Solución: Crear herramientas que nos permitas separar bien la información, conocer más el virus para poder llegar a formar la vacuna indicada, así como ir descubriendo más medidas de prevención de acuerdo con los casos que se han dado, así como llevar un control de estos.

Nombre de la Base de Datos:

Wine Reviews

Objetivo: Dada la alta demanda en este mercado, se quiere crear un modelo predictivo para identificar vinos, el primer paso en este viaje fue recopilar algunos datos para entrenar un modelo. Se planea usar el aprendizaje profundo para predecir la variedad de vinos. Esto nos puede ayudar a crear mejores contenidos e identificar errores, dar un mejor servicio a los consumidores y conocer más sobre los distintos tipos de vinos más y menos consumidos.

Planteamiento del Problema: Lograr crear alguna herramienta que pueda inspeccionar los distintos tipos de vinos, recopilar la información suficiente para poder tener bien estudiado el mercado y los distintos gustos de las personas. Se tomarán en cuenta también los costos y precios en los distintos lugares estudiados para hacer diferentes estudios y comparaciones.

Solución: Crear una herramienta que nos ayude a de alguna manera 'catar' vinos, poder llegar a la gente de una manera más profunda y conocer lo que realmente los atrae de los diferentes tipos de vinos, estudiar el mercado desde distintos ángulos, tomando en cuenta hacia qué publico nos estamos dirigiendo.

Nombre de la base de Datos:

Iris Species

Incluye tres especies de iris con 50 muestras cada una, así como algunas propiedades de cada flor. Una especie de flor es linealmente separable de las otras dos, pero las otras dos no son linealmente separables entre sí.

Objetivo: Conocer los distintos tipos de plantas a estudiar, sus características y diferencias entre sí, lugar donde crecen con mayor facilidad, especies en que se dividen y distintos datos más que nos ayuden a conocer mejor nuestro objeto de estudio.

Problema Planteado: Puede llegar a ser una base de datos muy grande, ya que siempre habrá especies que incluso ni siquiera conocemos aún, y esto es una de las cosas que lo hace interesante, podemos clasificar de acuerdo al lugar donde se originan, medidas y diferentes condiciones que necesitan para crecer y desarrollarse de una buena manera. También podríamos saber un poco más acerca de los usos que se les pueden dar, ya sean buenos o malos, el impacto que tiene en la naturaleza su existencia y beneficios que nos puede traer a los seres humanos.

Solución: Alguna herramienta que nos permita separar de manera práctica los distintos tipos de plantas (flores) para poder estudiarlas tanto por separado como en conjunto, relacionar los usos y el lugar donde crecen.

Nombre de la base de Datos:

Netflix Movies and TV Shows

Aquí se muestra todo un conjunto de datos que contiene programas de televisión y películas que están disponibles en la plataforma Netflix desde el 2019; en 2018, publicaron un informe que muestra que la cantidad de programas de televisión en Netflix casi se ha triplicado desde 2010. La cantidad de películas del servicio de transmisión ha disminuido en más de 2,000 títulos desde 2010, mientras que la cantidad de programas de televisión casi se ha triplicado.

El objetivo será conocer y deducir más cosas o incluso patrones de esta base de datos, podríamos conocer también qué tipo de públicos existen y cuáles son los más marcados que utilizan esta plataforma, ya que es algo muy común y que la mayoría usamos o por lo menos conocemos.

Problema planteado: Debemos comprender qué contenido está disponible en cada país, analizar contenidos similares y redes de actores o directores para poder comprender al público y saber qué contenido ofrecer, esto no es tarea sencilla ya que se habla de una plataforma que usan millones de personas en todo el mundo.

Solución: Crear una herramienta que nos ofrezca una clasificación, ya sea por medio del contenido o por países, para poder identificar patrones y buscar soluciones.