TIC

TALENTO TECH

Misión 1

INTELIGENCIA ARTIFICIAL

Explorador

Tema 4: Recolección de datos

UTraining Entrenamiento en Innovación y Tecnología | UNIVERSIDAD DE ANTIOQUIA | Universidad de Caldas | ubicua Technology

Campista, llegó el momento de retar tus conocimientos y que los pongas aprueba a través de los diferentes recursos que encontraras en este espacio como son: conceptos, ejemplos, herramientas, actividades prácticas y retos, los cuales te ayudaran alcanzar los objetivos trazados en el nivel explorador.

Recolección de Datos en el Ciclo de Vida de una Aplicación de Aprendizaje de Máquina

La recolección de datos es una fase fundamental en el ciclo de vida de una aplicación de aprendizaje de máquina, ya que la calidad y cantidad de los datos disponibles pueden determinar el éxito del modelo. A continuación, se describen los aspectos clave de esta etapa junto con ejemplos relevantes.

https://youtu.be/Qt-pN9CqJeo

1. Fuentes de datos

Las fuentes de datos son los orígenes de donde se obtiene la información necesaria para entrenar y evaluar los modelos de aprendizaje de máquina. Estas fuentes pueden ser variadas y dependen del problema específico que se está tratando de resolver.

Ejemplos de fuentes de datos

- *Datos Públicos:* Bases de datos gubernamentales, organizaciones internacionales, y repositorios abiertos.
- *Datos Privados:* Datos internos de la empresa, registros históricos, datos transaccionales.
- *Datos de Sensores:* Información recopilada por sensores en plantas de energía, dispositivos IoT, satélites.
- *Datos de Redes Sociales:* Información extraída de plataformas como Twitter, Facebook, Instagram.

Ejemplo:

Para un proyecto de democratización de la generación y consumo energético, las fuentes de datos pueden incluir registros de consumo de energía de hogares y empresas, datos de producción de plantas solares y eólicas, y datos meteorológicos.

2. Ejemplos de conjuntos de Datos Disponibles

Hay numerosos conjuntos de datos disponibles que pueden ser utilizados para proyectos de aprendizaje de máquina. Aquí algunos ejemplos relevantes:

Ejemplo 1:

Energía Renovable: Conjunto de datos de producción de energía solar y eólica (por ejemplo, el Open PV Project o el Global Wind Atlas).

Ejemplo 2:

Agricultura y Clima: Conjuntos de datos del Departamento de Agricultura de EE.UU. (USDA) y la Administración Nacional Oceánica y Atmosférica (NOAA).

Ejemplo 3:

Datos de Ciencia e Innovación: Conjuntos de datos de publicaciones científicas, patentes, y bases de datos de innovación (como la base de datos de la Oficina Europea de Patentes).

3. Tipos de bases de datos

Las bases de datos utilizadas para almacenar y gestionar los datos pueden variar según la estructura y el tipo de información. Aprendizaje de máquina. Este análisis incluye evaluar la disponibilidad y calidad de los datos, los recursos técnicos y humanos necesarios, y las limitaciones y riesgos potenciales.

Tipos de Bases de Datos

- *Bases de Datos Relacionales (SQL):* MySQL, PostgreSQL, Oracle.
- *Bases de Datos NoSQL:* MongoDB, Cassandra, Redis.
- *Data Warehouses:* Amazon Redshift, Google BigQuery.
- *Lakes de Datos:* Apache Hadoop, Microsoft Azure Data Lake.

Ejemplo:

Una base de datos relacional puede ser utilizada para almacenar datos estructurados de consumo energético, mientras que un lake de datos podría ser adecuado para almacenar grandes volúmenes de datos no estructurados de sensores y redes sociales.

4. Carga de Datos

La carga de datos implica importar y procesar los datos desde sus fuentes hasta el sistema donde se realizarán los análisis y entrenamientos de modelos.

Procesos de Carga de Datos

- *ETL (Extracción, Transformación, Carga):* Procesos que extraen datos de diversas fuentes, los transforman según las necesidades del análisis, y los cargan en un sistema de almacenamiento.
- *Stream Processing:* Procesamiento en tiempo real de flujos de datos utilizando herramientas como Apache Kafka y Apache Flink.

Ejemplo:

Para un proyecto de energía solar, los datos de producción pueden ser extraídos diariamente de sensores en los paneles solares, transformados para normalizar las unidades de medida, y cargados en una base de datos central para análisis.

5. Generación de Datos Sintéticos

Cuando los datos reales son insuficientes o difíciles de obtener, la generación de datos sintéticos puede ser una alternativa útil.

Métodos de Generación de Datos Sintéticos

- *Modelos Estadísticos:* Generar datos basados en distribuciones conocidas
- *Simulaciones:* *Uso de simulaciones computacionales para crear datos realistas.*
- *Algoritmos Generativos:* Redes Generativas Adversarias (GANs) para generar datos que imiten características de datos reales.

Ejemplo:

En un proyecto de predicción de consumo energético, se pueden generar datos sintéticos que simulen el comportamiento de consumo de nuevos tipos de dispositivos eléctricos o patrones de uso en diferentes condiciones climáticas.

6. Calidad y cantidad de Datos

La calidad y cantidad de los datos son factores cruciales que afectan el rendimiento del modelo.

Calidad de datos

- *Precisión y Exactitud:* Los datos deben ser correctos y representar con precisión la realidad.
- *Consistencia:* Los datos deben ser coherentes en todos los registros y fuentes.
- *Completitud:* No deben faltar datos esenciales.

Ejemplo:

Para asegurar la calidad de los datos en un proyecto de energías limpias, los registros de producción de energía solar deben ser precisos, consistentes y completos. La cantidad de datos debe ser suficiente para capturar variaciones diarias y estacionales en la producción de energía.

Material Complementario

Campista, en este espacio encontraras ayudas en diferentes formatos que pueden potenciar tu proceso de aprendizaje.

- ***Tipos de Datos en los Entornos de Big Data***
  https://youtu.be/dHM-kuxz4w4

- ***¿Dónde Encontrar DATOS para proyectos Data Science? ¿Qué es Kaggle?***
  https://youtu.be/NhHTWGIglRI

- ***CURSO de PYTHON con PANDAS Para Ciencia de Datos***
  https://youtu.be/2KCQQHpi2Qk

- ***Python for Data Science - Course for Beginners (Learn Python, Pandas, NumPy, Matplotlib)***
  https://youtu.be/LHBE6Q9XlzI