DATA FRAMES

Ramon Ceballos

21/1/2021

Formato de los datos en el siglo XXI

Toda la información se encuentra organizada de diferentes formas. Los datos se clasifican en:

- Datos no estructurados
- Datos semiestructurados
- Datos no estructurados

1. Datos no estructurados

La fuente de datos procede de un archivo de texto (txt) o binario (csv). No tiene una estructura definida. El formato clásico es el csv. Son muy flexibles pero complejos de manejar.

2. Datos semiestructurados

Tienen cierta estructura pero no cumplen una organización racional como puede ser una tabla o un gráfico.Es flexible también, teniendo algo de estructura.

Ejemplos populares: XML o JSON.

3. Datos estructurados

Es el más alto nivel de organización siendo fácil de tratar y predecir. Suele tener un tamaño máximo definido. Tienen reglas específicas a la hora de acceder a ellos. Es el más eficiente en cuanto a almacenamiento y eficiencia. Requiere un equipo detrás que se encargue de él.

Los ejemplos más comunes son MySQL o Apache ORC.

4. Tipos de documentos (Texto, csv, XML, JSON)

El documento de texto (txt) no tiene una estructura definida. Es cómodo para el uso del usuario.

Cada sistema operativo (Window, Linux) tiene su propia forma de definir el documento, por ello es complejo de tratar a posteriori por parte del desarrollador.

El documento csv es un archivo separado por comas. Es un archivo compacto que tiene una misma estructura. Los datos suelen ser inservibles ya que es dificil de interpretarlo. La estructura del archivo debe de ser respetada en cada una de las filas (mismas comas en cada una de ellas). Es bastante usada ya que se combina con excel... Es muy utilizado para crear un data frame y el análisis de datos.

Los datos XML es un lenguaje de marcas extensibles. Hya etiquetas () que abren una definición generando una estructura más pautada. En las etiquetas se identifica el dato y no tienen porque aparecer siempre las mismas etiquetas. Es público siendo open source. No es un formato muy amigable para análisis de datos.

El fichero JSON hace referencia a las siglas Java Script Object Notation (notación para objetos de JavaScript). Es un formato de archivo estandar y abierto, trasmitiendo conjunto de objetos y datos que consisten en pares de (clave):(valor o arrays) respectivamente (se llaman objetos serializables, por tanto, se pueden escribir en un fichero).

Ejemplo -> "Title":"The recipe"

Puede ser leído por cualquier lenguaje de programación, y eso es una ventaja. Actualmente es muy usado, desplazando a XML como formato de intercambio de datos en la red.

5. Bases de Datos: datos relacionales y no relacionales

Una base de datos es el formato ideal para ordenar, estructurar e intercambiar información. La gran mayoría de las bases de datos se manejas con un mismo lenguaje conocido como SQL (Structure Query Language).

Las bases de datos se dividen en dos grandes conjuntos: BBD relacionales y BBD no relacionales.

5.1. BBD relacionales

Constan de una serie de tablas relacionadas de diversos modos.

BBD relacionales se trata de una recopilación de elementos de datos con una serie de relaciones definidas entre ellos. Los elementos se organizan de forma conjunta en una tabla con cada una de las evidencias u observaciones en filas; y cada una de las características que consta en columnas.

Cada una de las tablas vienen a simbolizar las entidades u objetos que se quieren representar. Cada columna de la tabla guarda un determinado tipo de dato con una caracteristica concreta; y las filas de la tabla representan la recopilación de valores relacionados con una sola observación (un solo objeto o entidad). Cada fila de la tabla podría identificarse con un identificador único denominado clave primaria o principal (primary key o pk).

Las filas de varias tablas pueden relacionarse a través de las claves foráneas (foreign key o fk). Por tanto, una tabla tiene claves primarias y puede tener claves foráneas que referencien a las claves priamrias de otra tabla.

Cada tabla se podría identificar como un csv que se relacionan a traves de unas claves especiales que son las claves foráneas.

5.2. BBD no relacionales

Están en auge para el tema de análisis de datos, machine learning e inteligencia artificial.

Están orientadas a ficheros o documentos para poder almacenar y recuperar datos que no sean necesariamente rectangulares (tablas de datos).

En general, las bases de datos no SQL son muchas más flexibles y escalables. Permite almacenar los datos de una forma no tan rígida como las BBD relacionales.

Trabajar con este tipo de BBDD es bastante pesado, porque estamos hablando de tecnología novedosa y uno debe de mantenerse al día con las actualizaciones. Permiten adicionar datos en una base dedatos previa sin tener que redefinir ésta, permite procesar gran volumen de datos sin estructura o semiestructurados. Permite almacenar datos en la nube. En este curso no le vamos a prestar demasiada atención.

Ejemplos: SQLServer, MySQL, Hadoop, Hive,etc

5.3 Lenguaje SQL

El lenguaje SQL es el lenguaje de comunicación para las tablas de una BBDD. Permite articular que información queremos obtener de una colección de tablas de la BBDD.

Los diagramas entidad-relación (diagrama de estructura de relaciones) son muy importantes para saber como se relacionan las tablas entre sí. Las entidades (cada una de las tablas) se pueden pintar como puntos, cículos, polígonos y óvalos (es un equivalente gramatical del sustantivo); y los atributos de las entidades () son cada una de las columnas de que consta la tabla o entidad. Las relaciones entre las distintas entidades (equivalente a verbos) se definen en función del conjunto de entidades que se asocian entre sí. La figura de las relaciones aparece en la carpeta.

Importante saber un poco de SQL para desarrollarse como data science. Leer con los informes de la web.