Visualización de datos

Introducción a la preparación de datos

Julià Minguillón

septiembre 2024

Índex

¿Qué son los datos?

¿Qué son los datos?

Representación simbólica de una entidad que describe algún aspecto o atributo de la misma

Un dato que responde una pregunta es **información**

Así, **42** no es más que un dato, pero cuando responde a la pregunta "¿Qué temperatura tiene el paciente?" se convierte en información

Los datos necesitan contexto, unidades, precisión, formato, etc. para poder ser útiles (ser analizados, visualizados)

Datos, información, conocimiento, sabiduría

Dato: 42

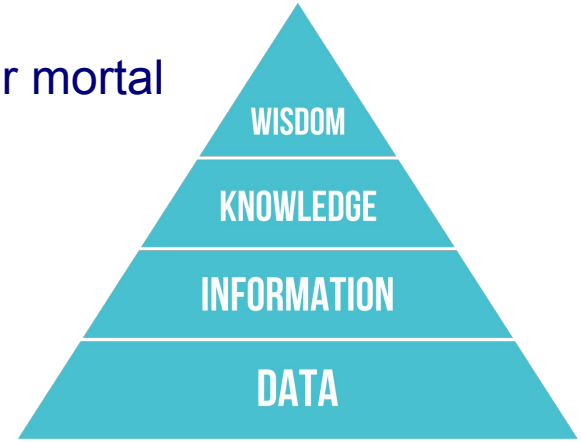
Información: temperatura del paciente en grados Celsius

Conocimiento: una fiebre de 42°C o más puede ser mortal

Sabiduría: no hay que dejar que la temperatura del paciente alcance ese valor

⇓

Usar el conocimiento para el bien común



¿Qué tipos de datos hay?

Simples: describen un solo elemento, aspecto, …
Estructurados: son una combinación de diferentes datos simples o estructurados, organizados (listas, matrices, …)

- Un color RGB es una tripleta de datos (R, G y B)
- Una imagen es una matriz 2D de píxeles RGB

Simbólicos: cierto/falso (booleanos), categorías (p.e. día de la semana)
Numéricos: enteros, reales, complejos
Texto: caracteres, cadenas, … (líneas, párrafos, …)

Metadatos: datos sobre los datos

En la práctica

Datos binarios: valores True / False (a veces 1 / 0), permiten operaciones lógicas (AND, OR, NOT)
Datos numéricos: describen cantidades, permiten operaciones matemáticas y comparaciones, pueden ser enteros o con decimales
Datos categóricos: describen aspectos, no permiten operaciones matemáticas

- **Ordinales**: existe un orden, permiten comparaciones
- **Nominales**: no existe un orden

Fecha / hora: permiten ciertas operaciones (p.e. día siguiente, diferencia) y comparaciones (antes de / después de)

Problemas típicos

***Outliers*: son valores atípicos que no concuerdan con la mayoría del resto de datos, pueden ser errores puntuales de registro o bien causados por un fenómeno inusual**

Missing data: datos que no han sido registrados y quedan en blanco o bien con valores tipo N/A, NA, etc.

Registros duplicados: en algunos casos hay un campo que hace de identificador único, pero aparecen dos o más registros con el mismo

Errores de formato: típicos en campos tipo fecha u otros donde hay una estructura prefijada que no se respeta

Múltiples categorías: cadenas de texto diferentes usadas como si fueran el mismo valor

Ejemplo: datos de usuarios

12345678	H	32	Barcelona	1-9-2012	Muy bien
13572468	M	128	Madrid	15/12/2010	Mal
24681357	M		Barcelona	01/08/2020	Regular
14235867	H	22		15/7/2021	Muy bien
14.235.867	H	22	BCN	15/7/2021	Bien

Problemas típicos

Outliers

Diferentes formatos

Duplicados / inconsistencias

Valores ausentes (*missing data*)

Diferentes categorías equivalentes

Garbage In, Garbage Out (GIGO)

GIGO: si los datos que alimentan un proceso son basura, el resultado será también basura

Es necesaria una **preparación de los datos** para poder analizarlos y visualizarlos posteriormente, eliminando todos los problemas típicos presentes

Objetivo: disponer de los datos necesarios en un formato estructurado para facilitar su análisis y visualización

Datos estructurados

Datos estructurados

Tabulares / matriciales: formato más sencillo para representar datos

- Fila: registro, individuo, elemento, …
- Columna: atributo / variable

Casi todo se puede representar en forma tabular con una o más tablas

Grafos: relaciones entre elementos

- Nodos: elementos (tabla de nodos)
- Aristas / arcos: relación entre elementos (tabla de relaciones)

Otros: datos geoespaciales, imágenes, audio, vídeo, etc.

El concepto de *tidy data*

***Tidy data*: datos “ordenados” en tablas sin problemas**

- Cada columna es una variable (atributo, campo)
- Cada fila es una observación (registro, elemento, individuo, …)
- Cada celda tiene un solo valor (dato)

Esto facilita las operaciones básicas con los datos

- Filtrado / selección de filas y columnas
- Fusión de dos tablas
- Pivote entre formatos *long* y *wide*
- Agregados

Ejemplos

A

Untidy Data

species	habitat	weight	length	latitude/longitude	date
Alligator mississippiensis	swamp	431 lb	4 ft 2	29.531,-82.184	Sept 15, 2015
Puma concolor	forest	125 lb	2.2m	29.125,-81.682	08/10/2015
Ursus americanus	forest	88 kg	133 cm	N29°7'30"/W81°40'55.2"	07-13-2015

B

Tidy Data

meta-data

data

species_code	date	station_code	weight_kg	length_cm
TSN 551771	2015-09-15	1	196	127
TSN 55247	2015-08-10	2	57	220
TSN 180544	2015-07-13	2	88	133

station_code	habitat	latitude	longitude
1	swamp	29.531	-82.184
2	forest	29.125	-81.682

species_code	class	genus	species
TSN 551771	Reptilia	Alligator	mississippiensis
TSN 55247	Mammalia	Puma	concolor
TSN 180544	Mammalia	Ursus	americanus

wide

id	x	y	z
1	a	c	e
2	b	d	f

long

id	key	val
1	x	a
2	x	b
1	y	c
2	y	d
1	z	e
2	z	f

Datos no estructurados o semi-estructurados

Texto: aunque forme parte de una entidad con cierta estructura (libro, artículo, …), el texto se considera datos no estructurados

Normalmente hay un proceso previo para extraer los datos que luego serán usados en forma de tabla:

- Contar frecuencias de caracteres, n-gramas, palabras
- Uso de ciertas palabras clave, conectores, gramática
- Análisis de contenido (estilo, nombres, personajes, …)

Analizar y visualizar texto es un reto muy complejo

Ejemplo de análisis de texto

Biografías en Wikipedia: mediante la API o los dumps de Wikipedia es posible acceder al contenido de una o más páginas de una o más categorías (biografías)
https://www.mediawiki.org/wiki/API:Categorymembers

Un estudio con la Wikipedia inglesa demostró que las palabras usadas en biografías de mujeres correspondían más a la familia y las relaciones (“hija de”, “esposa de”, …) y que tenían más enlaces a biografías de hombres

Referencia
https://link.springer.com/content/pdf/10.1140/epjds/s13688-016-0066-4.pdf

Obtención de datos

Datos en abierto (*Open Data*)

Definición: datos que pueden ser localizados, accedidos, usados, reusados y redistribuidos sin “ninguna” restricción tecnológica o jurídica

Posibles restricciones (en la práctica):

- Registro para acceder a los datos
- Coste económico para acceder a los datos: abierto ≠ gratis
- Limitaciones en la redistribución
- Dificultades en el uso ⇒ formato de los datos, necesidad de usar herramientas propietarias, …

El modelo de cinco estrellas de *Open Data*

Primer nivel: datos disponibles (en cualquier formato), p.e. en PDF

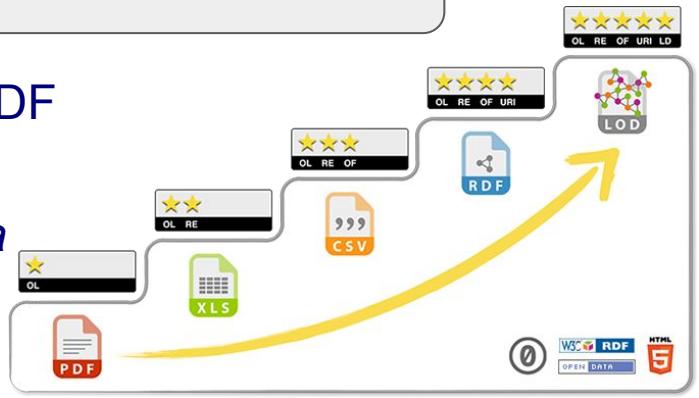
Segundo nivel: disponibles en un formato estructurado, p.e. XLSX

Tercer nivel: en un formato estructurado abierto, p.e. CSV, JSON

mínimo recomendado

Cuarto nivel: permiten acceso vía URI, p.e. RDF

Quinto nivel: incluyen contexto vía *linked data*



Fuente: https://5stardata.info/es/

Fuentes de datos en abierto

Repositorios genéricos: data World Bank, Kaggle, UCI ML, …

Fuentes oficiales: INE, IDESCAT, EUROSTAT, UN SDG, …

Bases de datos: genómicas, imágenes de satélite, citas bibliográficas, …

Redes sociales: Twitter (limitado), Wikipedia, Instagram, …

Sensores: AEMET, datos de la Bolsa, imágenes CCTV, tráfico marítimo, …

Muchas permiten un **acceso estático** y algunas también **dinámico**

Acceso estático

Son “fotografías” del estado de una colección de datos en un momento dado

Ejemplo: volcados de Wikipedia en un día concreto
https://dumps.wikimedia.org/eswiki//latest/

Suelen ser uno o más ficheros en formatos CSV, JSON, XML u otros más específicos del ámbito

No siempre es posible especificar qué registros queremos que aparezcan en los ficheros disponibles

Acceso dinámico

Los datos se acceden mediante una **API** que responde a una petición (*query*) de acuerdo a unos parámetros

Los **parámetros** determinan qué datos se quieren recuperar y en qué formato

Normalmente, hay un límite de peticiones y/o de resultados obtenidos por petición

El resultado suele ser un fichero CSV, JSON o XML con los datos requeridos

Ejemplo de *query*

API de flickr: nos permite acceder a todos los contenidos de flickr (imágenes, usuarios, grupos, etiquetas, …)

Ejemplo: recuperar imágenes geolocalizadas en la Sagrada Familia en Barcelona desde 2022

Query: desde https://www.flickr.com/services/api/explore/flickr.photos.search solicitamos (rellenando los campos del formulario)
https://www.flickr.com/services/rest/?method=flickr.photos.search&api_key=8857a51562ecac8bb8549f540f2ae26f&tags=Sagrada+Familia&min_upload_date=2022%2F01%2F01&has_geo=yes&lat=41.403638&lon=+2.174388&radius=1&format=rest

Resultado (con una api_key válida)

```
<?xml version="1.0" encoding="utf-8" ?>
<rsp stat="ok">
  <photos page="1" pages="2" perpage="100" total="167">
    <photo id="52512618299" owner="36988894@N02" secret="09907659cc" server="65535" farm="66" title="Barcelona - Sagrada Familia" ispublic="1" isfriend="0" isfamily="0" />
    <photo id="52512349821" owner="36988894@N02" secret="7f2bb78def" server="65535" farm="66" title="Barcelona - Sagrada Familia" ispublic="1" isfriend="0" isfamily="0" />
    <photo id="52511874012" owner="36988894@N02" secret="65764255f3" server="65535" farm="66" title="Barcelona - Sagrada Familia" ispublic="1" isfriend="0" isfamily="0" />
    …
</photos>
</rsp>
```

Nos dice que hay 167 fotos en total y nos devuelve 100 en la primera página de resultados (se necesita una segunda *query* para el resto)

Podemos usar https://www.flickr.com/services/api/explore/flickr.photos.getInfo para obtener información de cada foto y acceder a la misma vía su URL a partir del *photo id* devuelto por la consulta anterior

Preparación de datos

El concepto de *data wrangling*

***Data wrangling*: se trata de preparar los datos obtenidos en "crudo" para poder "cocinarlos" después**

- Desechar los datos inservibles
- Seleccionar los necesarios
- Limpiarlos
- Trocearlos
- Mezclarlos
- Enriquecerlos

Fase crucial del **ciclo de vida** de los datos (preprocesamiento)
http://datascience.recursos.uoc.edu/es/ciclo-de-vida-de-los-datos/

Operaciones básicas

Filtrado (SELECT): selección de las filas y/o columnas de acuerdo a ciertos criterios

Selección de filas: elementos que cumplen una condición, y también

- Descartar filas con muchos datos ausentes
- Muestrear una tabla demasiado grande

Selección de columnas: atributos que se consideran relevantes para el objetivo a conseguir, y también

- Reducción de la dimensionalidad
- Reordenar columnas

Operaciones básicas

Fusión (JOIN): enriquecer una tabla con datos provenientes de una segunda, ambas tablas comparten un campo clave único

Ejemplo: al fichero de notas (*usuario*, semestre, asignatura, nota) le queremos añadir algunos datos del fichero de perfiles (*usuario*, fecha de nacimiento, género, …)

USUARIO	SEMESTRE	ASIGNATURA	NOTA
12345678	2021/2	05.554	NO
12345678	2021/2	05.562	A
24681357	2021/2	05.570	A
…	…	…	…

USUARIO	FECHANAC	GÉNERO
12345678	23/06/1981	H
24681357	01/12/1977	M
…	…	…

Operaciones básicas

Pivote: en el formato ***wide*** hay una columna para cada valor posible de una variable (p.e. gasto mensual)

USUARIO	ENERO	FEBRERO	…	DICIEMBRE
12345678	125	120		200
13572468	150	220		250
…				

Nuevas variables clave, valor

USUARIO	MES	VALOR
12345678	ENERO	125
12345678	FEBRERO	120
…	…	…
12345678	DICIEMBRE	200
13572468	ENERO	150
…	…	…

El formato ***long*** suele ser más adecuado para muchas operaciones con los datos ⇒ hay una fila para cada valor posible de la columna clave que se usa como pivote

Operaciones básicas

Agregados (GROUP): a partir de una o más filas que satisfacen un criterio (forman un grupo), aplicar una operación sobre alguna variable

Operaciones típicas

- Contar el número de apariciones / apariciones únicas
- Descriptores estadísticos: promedios, máximos, mínimos

Ejemplo: a partir de las notas de un estudiante de cada semestre, podemos generar un nuevo conjunto con la nota media del expediente, el número de semestres en activo y el total de asignaturas matriculadas y aprobadas por estudiante

Variables *dummy*

Objetivo: convertir una variable categórica de N valores en N-1 variables binarias (0/1, presencia/ausencia)

Ejemplo: si tenemos una variable que indica el tipo de dispositivo que usa un estudiante con los 5 valores *ordenador*, *laptop*, *tablet*, *móvil*, *otro*, podemos crear 4 variables *dummy* llamadas *usa_ordenador*, *usa_laptop*, *usa_tablet*, *usa_movil*, que valdrán 0 si ese estudiante no usa ese tipo de dispositivo y 1 si lo hace

No hace falta codificar los N valores, el valor por defecto ausente se suele asignar a la categoría más popular o al valor de referencia

Extracción de características

Objetivo: añadir valor al conjunto de datos mediante el cálculo de nuevas variables a partir de las existentes

Ejemplo: a partir del peso y la altura se puede calcular el índice de masa corporal (https://es.wikipedia.org/wiki/Índice_de_masa_corporal)
Ejemplo: a partir de una colección de ítems de respuesta de una encuesta, calcular los factores más importantes (mediante PCA)
Ejemplo: a partir de los logs de conexión al campus virtual, calcular el número de sesiones de cada usuario

Se usan fórmulas, métodos estadísticos o de minería de datos

Descriptores estadísticos

Objetivo: obtener información sobre la naturaleza de los datos

Moda, Media, Mediana: alrededor de qué valor o valores se distribuyen los datos
Varianza, Kurtosis: cómo de dispersos y sesgados están los datos
Box-plot: visualización de los cuartiles y otros descriptores importantes
Histograma: distribución de los valores posibles de una variable
Correlación: medida de asociación entre dos variables

Atención: diferentes conjuntos de datos pueden tener los mismos descriptores! http://datascience.recursos.uoc.edu/es/drawmydata/

Herramientas

Manipulación de PDFs

Tabula: extracción de tablas de ficheros PDF
https://tabula.technology/

PDFsam Basic: fusión, extracción, rotación, … de documentos PDF
https://pdfsam.org/pdfsam-basic/

XpdfReader: conversión de PDF a otros formatos (p.e. texto)
https://www.xpdfreader.com/about.html

Conversión de formatos

DenCode: codificación y decodificación de texto, números, fechas, etc. en diferentes formatos
https://dencode.com/

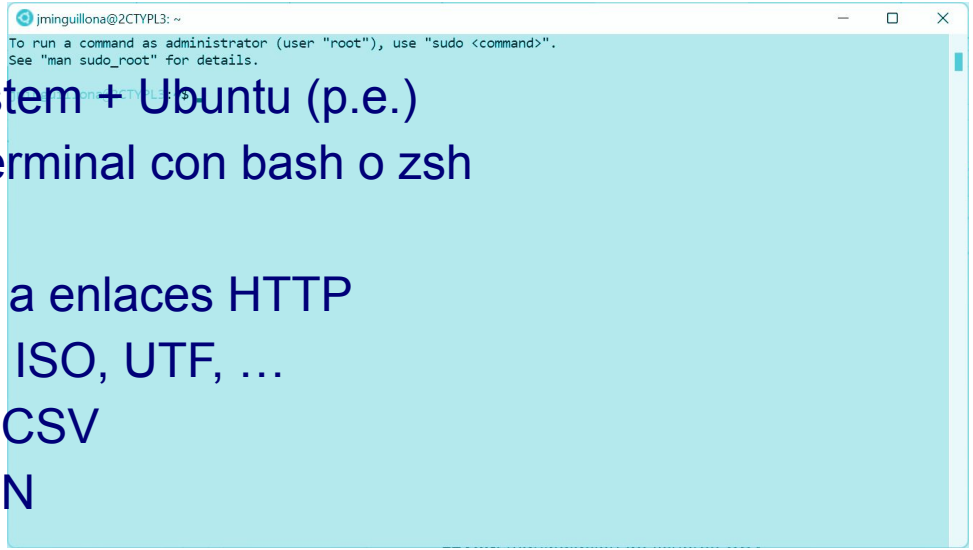
Mr. Data Converter: herramienta online para convertir formatos
http://shancarter.github.io/mr-data-converter/

Web scraping

Scrapy / Beatifulsoup: librerías en Python para la extracción de contenidos de páginas web
https://scrapy.org/
https://pypi.org/project/beautifulsoup4/
rvest: librería en R para la captura y análisis de páginas HTML
https://rvest.tidyverse.org/
instant data: extensión para el navegador para capturar datos de páginas web
https://webrobots.io/instantdata/
ParseHub: aplicación para la extracción de páginas HTML
https://www.parsehub.com/

Herramientas de la línea de comandos

Desde la línea de comandos del sistema operativo es posible realizar algunas operaciones básicas



Windows: desde el Linux subsystem + Ubuntu (p.e.)
MacOS / **GNU/Linux**: desde el terminal con bash o zsh

cURL: herramienta para acceder a enlaces HTTP
iconv: conversión entre formatos ISO, UTF, …
csvkit: manipulación de ficheros CSV
jq: manipulación de ficheros JSON

R + RStudio + CRAN

R: lenguaje de programación orientado a la manipulación, análisis y visualización de datos
https://www.r-project.org/about.html

RStudio: entorno de desarrollo en R
https://posit.co/products/open-source/rstudio/

CRAN: archivo de paquetes para R orientados a resolver problemas específicos
https://cran.r-project.org/web/views/

Tidyverse

Tidyverse: colección de paquetes para R orientados a la manipulación de datos
https://www.tidyverse.org/packages/

dplyr: operaciones básicas con tablas
tidyr: mantenimiento de datos “ordenados” (*tidy data*)
ggplot2: generación de gráficos
…

Python

Python: lenguaje de programación de propósito general muy usado en ciencia de datos
https://www.python.org/

dataprep: módulo para la preparación de datos en Python
https://dataprep.ai/

pandas: librería completa para la manipulación, análisis y visualización de datos en Python
https://pandas.pydata.org/

Otras herramientas

Tableau Prep: herramienta visual para la preparación de datos
https://www.tableau.com/es-es/trial/tableau-prep
DataWrangler: herramienta online para limpiar y transformar datos
http://vis.stanford.edu/wrangler/app/
OpenRefine: herramienta avanzada para la manipulación y extracción de datos mediante filtros y facetas de ficheros CSV, JSON, RDF, XML, …
https://openrefine.org/
TAGS: captura de tweets mediante Google Sheets
https://tags.hawksey.info/ => desde el cambio de Twitter a X funciona parcialmente
Power Query: herramienta integrada en Excel para manipular datos
https://learn.microsoft.com/es-es/power-query/power-query-what-is-power-query

Para saber más

Para saber más

Pirámide DIKW: https://en.wikipedia.org/wiki/DIKW_pyramid

Tipos de datos: https://en.wikipedia.org/wiki/Data_type

Open Data: https://opendatahandbook.org/guide/es/what-is-open-data/

Fuentes de datos: https://datacatalog.worldbank.org/

Búsqueda de datos: https://datasetsearch.research.google.com/

Lista de APIs disponibles: https://github.com/public-apis/public-apis

El lenguaje SQL: http://sqlfiddle.com/

Línea de comandos: https://link.springer.com/book/10.1007/978-1-4842-0121-3

Descriptores estadísticos:
https://learning.oreilly.com/library/view/statistics-in-a/9781449361129/ch04.html

EDA: https://en.wikipedia.org/wiki/Exploratory_data_analysis

Universitat Oberta
de Catalunya

UOC.universitat
@UOCuniversitat
UOCuniversitat