2021-09-21

Anton-Bravo, Adolfo

Contents

1 Contenidos

1.1 Datos

Hemos repasado los términos de periodismo de datos a propósito de lo que haremos después.

- Periodismo de datos alude al uso de datos pero no hay que olvidar que estos datos son registros electrónicos

- El hecho de llamarlo "periodismo guiado por datos" o *data driven journalism* no debe menospreciar que lo importante es el periodismo.

- Pero hay que saber trabajar con datos como una parte más del trabajo periodístico.

- El *Computer Assisted Reporting* o periodismo asistido por ordenadores lleva funcionando en EE.UU. desde los 1950.

1.2 Tipos de formatos de datos

Aunque no hemos empezado por aquí, lo hago así para que se entienda mejor por parte de quien lo lea.

En este caso no hablamos de las bases de datos y de SQL sino de los tres tipos de formatos de datos de ficheros que nos vamos a encontrar habitualmente:

1. **SV* o valores separados por cualquier valor.

2. *JSON* o *JavaScript Object Notation*, notación de objetos *JS*.

3. *XML* o *eXtensible Markup Language*, lenguaje de marcas extensible.

1.2.1 XML

Los ficheros XML no los vamos a ver porque:

- Son más complicados de leer.

- Por tanto, es más complicado trabajar con ellos

1.2.2 JSON

- Son los ficheros que mejor funcionan con aplicaciones web.
- Utilizan la sintaxis de *JS*.
- Permiten más complejidad que los **SV*, por eso son más complejos de leer.

1.2.3 **SV*

- Normalmente serán llamados con terminación *csv* incluso aunque no utilicen comas para separar los valores.
- Son los más sencillos.
- Pero también los menos estandarizados.
- Los valores separados por comas se visualizan como una tabla simple con filas y columnas.
- La mayoría de los recursos disponibles en los catálogos de Datos Abiertos se encuentran en formato CSV.
- Elportaldedatoseuropeodisponedemásde120milconjuntosdedatosenformatoCSV,siendoelform

1.3 Aprender a partir de una tabla de datos

- A partir del listado de la clase aprendemos algunas cosas de las tablas.
- La tabla es una representación visual de un **SV*, un archivo de valores separados por comas.

- La tabla la leemos de izquierda a derecha y de arriba a abajo.

- La lectura horizontal se corresponde a las filas y la vertical a las columnas.

- Las filas son las "líneas" del archivo.

- A veces, aunque no siempre, la primera línea es la **cabecera** de la tabla e indica qué información tiene cada columna.

- La información de la columna está relacionada con el **tipo de datos** que tiene la tabla.

- Cada intersección de fila y columna es una celda.

- Cada variable es una columna

- datos.gob.es ha [[`https://datos.gob.es/sites/default/files/doc/file/guia_csv_vf.pdf`

][publicado]] una guía muy completa, también en formato chuleta.

1.4 Tipos de datos

1.4.1 Numéricos

- Cuando una celda tiene números es probable que esos datos sean **numéricos**.

- Pero no siempre ocurre ya que solo son considerados *numéricos* si queremos realizar operaciones matemáticas con ellos.

- En nuestro caso, estos números corresponden a un identificador, que en este caso está compuesto por números.

- Por tanto, en este caso estos números no son datos *numéricos* sino *string*, "cadena de caracteres" o *literales*.

- Algunas aplicaciones reconocen automáticamente o pueden hacerlo los tipos de datos para realizar operaciones específicas con ellos.

- Esto suele mostrarse visualmente de alguna manera destacada, por ejemplo, poniendo esos datos en color verde.

Tipos de datos numéricos

integer números enteros, sin decimales. Atención si tienen el separador de millar porque es distinto en español (punto) que en inglés (coma). Algunos programas lo entienden según tu codificación del programa pero otras veces hay que indicarlo.

decimal números con decimales pero –explicación corta– pocos decimales y siempre el mismo número de decimales. Por cómo son tratados, son más lentos de procesar que los `float`.

float or double números con decimales pero que pueden tener muchos decimales y/o variable en su longitud. Por cómo son tratados son más rápidos de procesar que los `decimal`

date or datetime la forma más estándar suele ser la que sigue el esquema `YYYY-MM-DD`, donde `Y` significa `Year`, y al ser cuatro tienen que ponerse cuatro cifras; `M` significa `Month`, y al ser dos tienen que ponerse dos cifras; y `D` significa `Day` y al ser dos tienen que ponerse dos cifras. Fíjese que en este tipo de datos numérico se utiliza un guión para separar las unidades temporales, aunque hay veces que se separan con `/`, no es lo más habitual. Hay veces que se incluye también la hora `time`, a continuación de la fecha, o bien separada con una `T` de `Time` o, simplemente, con otro guión, en la forma `HH:MM:SS`: `2021-09-21-14:30` o

`2021-09-21+14:30`. Se suelen poner horas y minutos si no se necesitan los segundos, pero puede haber segundos e incluso décimas de segundos: `2021-09-21T14:30:00.5`. También se puede indicar la zona temporal añadiendo una `Z` al final que indica que se está en horario `UTC` (*Universal Time Coordinated*... en realidad no está en inglés exactamente). Madrid está en UTC+2 en horario de verano y UTC+1 en horario de invierno por lo que, la fecha anterior se escribiría en Canarias así: `2021-09-21T14:30:00.5Z` pero en Madrid sería `=2021-09-21T14:30:00.5+2=`. El mapa con los husos horarios lo tenéis en la Wikipedia. El tema de las fechas se ha especificado tanto quizás porque ha generado unos cuantos problemas informáticos. Véase, por ejemplo, el problema del año 2000 que fue bastante comentado: CCFN TV, NatGeo, The Science Elf. Pero, ¿puede volver a pasar? ¡Sí! Ya tenemos el Year 2038 Problem por el registro de hora en 32 bits. La solución pasa por hacerla en 64 bits. Ver formatos de fecha y hora

period Algunas veces (duration data type de XML Schema) se utiliza también el tipo de dato de tiempo periódico que obedece al periodo de la muestra del dato, por ejemplo:

- `P` al inicio indica que se trata de un dato periódico.
- `nY` indica el número de años.
- `nM` indica el número de meses.
- `nD` indica el número de días.
- `T` indica el comienzo de horas, minutos o segundos, según vaya `nH`, `nM` o `nS`
- En este tipo de dato se pueden dar valores negativos para indicar mediciones aproximadas. Por ejemplo, si es `-P10D` indica un periodo menor a diez días.

1.4.2 *Strings*

- Se denomina *strings*, cadena de caracteres o literales al texto normal.

1.4.3 Booleanos

- Representan dos valores de una lógica binaria.

- "Verdadero o Falso", "True or False", "Sí o No", "0 o 1", etc.

- El nombre se debe a George Boole, "desarrolló un sistema de reglas que le permitían expresar, manipular y simplificar problemas lógicos y filosóficos cuyos argumentos admiten dos estados (verdadero o falso) por procedimientos matemáticos."

1.5 Repaso de infraestructura

- Hay que instalarse OpenRefine, la navaja suiza de la limpieza de datos. Si no podéis lo hacemos el próximo día.

- Instalación de la terminal...

1.5.1 XCode

- Hay gente con MacOSX que no ha podido instalar XCode porque le sale un aviso de que no tiene espacio en disco.

- Este tipo de mensajes están bien, son normales, las aplicaciones ponen mensajes... pero eso no tiene que frenarnos para nuestros propósitos. Hemos de preguntarnos qué podemos hacer.

- Una opción es mirar el tamaño de nuestro disco duro. Se puede hacer por aplicaciones gráficas o bien con el comando `df`:

language=bash,label= ,caption= ,captionpos=b,numbers=none df -h

Si no tenemos espacio, debemos buscar la forma de tenerlo.
Si tenemos, podemos buscar ayuda:

- A alguien que sepa.
- A otras personas, en el foro de la clase.
- A mí.
- A tu buscador favorito.

Si usamos duckduckgo para eso con una búsqueda tipo xcode fail install disk space y hemos hecho una pregunta inteligente:

- `https://stackoverflow.com/questions/53432700/xcode-on-mac-app-store-cant-install 55518395`
- `https://discussions.apple.com/thread/8622103?answerId=250008933022# 250008933022`

Si eso no nos ayuda, seguir buscándolo.

1.5.2 Cygwin

Lo vemos el próximo día

1.5.3 Tablet Windows 10

No parece tener ningún problema para instalar programas Windows.

1.6 Periodismo y Visualización de datos

- Se habla de periodismo y visualización de datos porque entendemos que hay una línea argumental entre ambos conceptos.

- El periodismo de datos usa la visualización de datos tanto en la etapa de análisis como en la de presentación de resultados.

- A la vez son términos que no se han definido por completo. ¿Se puede hablar de periodismo y visualización sin análisis? No, pero, la visualización remite también al análisis de datos.

1.7 Enlaces

- He encontrado este interesante artículo donde hablan de periodismo de precisión y le trasladan unas preguntas al propio Philip Meyer, que las responde ampliamente.

- También me gustaría que vierais el vídeo, un corte de una entrevista a Philip Meyer donde habla del *Harvard Data Text*

2 Pruebas

- Cuando hablamos de periodismo o visualización de datos, ¿a qué datos nos referimos? Razona la respuesta.

- ¿Qué tipos de formatos de datos hay? ¿Que similitudes y diferencias tienen?

- ¿Que tipo de dato de fecha elegirías para tus archivos? Razona tu respuesta.