Actividad 1: Limpieza de *dataset*

Objetivos

Con esta actividad vas a reconocer una de las tareas más importantes y que más a menudo realiza un científico de datos: la limpieza del *dataset.* Además, trabajarás la conversión entre formatos y la importación de documentos en una base de datos, lo que supone un proceso de captura y almacenamiento de los datos.

Descripción de la actividad y pautas de elaboración

En el fichero de datos adjunto (fichero: *data\_act\_01.cvs*) hay al menos seis errores que se han introducido intencionadamente y que requieren de cierta limpieza y homogeneización. Se pide:

1. Para cada uno de los elementos susceptibles de ser limpiados, indicar:
   * Descripción del problema, señalando claramente:
     + Qué problema es.
     + Cómo se ha detectado.
     + Por qué crees que se trata de un problema.
     + Un ejemplo completo extraído del propio *dataset.*
   * Forma de solucionarlo. Ten en cuenta que un mismo problema puede tener varias soluciones diferentes, así que deberás escoger la que consideres más adecuada. Indica, en tu solución, una justificación de por qué la solución es adecuada. Es decir, de por qué el estado final del *dataset* tras aplicar la solución es mejor que el inicial. En la solución de este primer apartado, ten en cuenta que cada problema detectado puede afectar a una o varias filas, por lo que en realidad no se está buscando un problema puntual, sino un tipo de problema.
2. Propón un formato JSON equivalente al CSV proporcionado en el que cada documento JSON contenga toda la información contenida en la fila correspondiente del documento CSV. Aprovecha esta transformación de formato para eliminar los campos redundantes.
3. Indica una metodología, clara y bien definida que permita afirmar que este catálogo de datos esté limpio y listo para su análisis. La metodología debe ser un listado de pasos fáciles de seguir, no ambiguos. Se valorará lo completo de la metodología y la claridad de aplicación de esta.
4. Propón al menos dos mejoras en el conjunto de datos. Ten en cuenta que estas mejoras pueden afectar al procedimiento de captura de datos o a la estructura de la base de datos, así que debes explicar tanto la mejora como el procedimiento para llevarla a cabo.

Entrega y extensión

Entrega un informe con los elementos habituales (portada, índice, nombre, fecha, numeración de páginas, etc.) en que des respuesta a los cuatro apartados planteados anteriormente. Debes responder cada uno de los apartados en una sección diferente. El fichero JSON debe acompañar al informe propuesto y no debe contener ningún error de formato (validar previamente el formato JSON del fichero).

El informe entregado no debe exceder las 6 páginas, sin contar portada e índices. Fuente utilizada Georgia 11 e interlineado 1,5.

Al finalizar la actividad, **harás la entrega del informe final** a través del apartado «Envío de actividades» del aula virtual. El informe y el fichero JSON a entregar deben ir nombrados como se muestra a continuación.

* APELLIDO1\_APELLIDO2\_NOMBRE\_actividad\_1.pdf
* APELLIDO1\_APELLIDO2\_NOMBRE\_actividad\_1.json
* Alternativa en ZIP: APELLIDO1\_APELLIDO2\_NOMBRE\_actividad\_1.zip

No uses tildes, apóstrofes ni ningún otro carácter que pudiera resultar conflictivo.

Rúbrica:

|  |  |  |  |
| --- | --- | --- | --- |
| Actividad 1: Limpieza de *dataset* | Descripción | Puntuación máxima  (puntos) | Peso  % |
| Criterio 1 | Apartado 1:  Para cada uno de los elementos limpiados:  -La descripción del problema es correcta.  -La solución aportada es correcta, y es replicable. | 6 | 60 % |
| Criterio 2 | Apartado 2:  -El formato propuesto JSON es correcto.  -El formato propuesto contiene la misma información que el CSV de origen.  -Se han eliminado los campos redundantes. | 1 | 10 % |
| Criterio 3 | Apartado 3:  - La metodología es adecuada.  - La metodología está explicada a través de pasos fáciles de seguir. | 1,5 | 15 % |
| Criterio 4 | Apartado 4:  - Las mejoras propuestas son pertinentes.  - Las mejoras propuestas están bien explicadas.  - La implantación de las mejoras propuestas está bien explicada. | 1,5 | 15 % |
|  |  | **10** | **100 %** |