UNIVERSITAT POLITÈCNICA DE VALÈNCIA

etsinf

Proyectos II, integración y preparación de datos

Primera presentación:

Captura de datos

Introducción

Antes de cada HITO de presentación, debemos rellenar estas fichas y presentarlas a través de PoliformaT, en la tarea que indiquen los profesores. Cada equipo de trabajo presenta las mismas fichas. Sólo será necesario que las suba uno de los componentes del equipo.

Nombres y apellidos de los autores:

Javier	Luque Sáiz
Daniel	Garijo Abarca
Pablo	Parrilla Cañadas
José	Valero Sanchis
Qilu Diana	Wu

1. Las Fichas de Configuración

Una vez hayamos decidido el proyecto en el que vamos a trabajar, debemos rellenar el Alcance preliminar del proyecto (apartados 1.1.).

Una vez definido el Alcance, desglosaremos el trabajo de esta primera etapa en:

- Localización de las fuentes.
- Técnicas de obtención de los datos y extracción. Por ejemplo:
    - Descarga de ficheros .csv
    - Descarga desde una URL
    - Lectura de tablas incrustadas en HTML

  - Conversión de .JSON
  - Conversión de .XML
  - Recoger datos de Twitter y limpieza sobre expresiones regulares
  - Recoger datos de Google y limpieza sobre expresiones regulares
  - Web scraping
  - Descargas en tiempo real durante varios ciclos

- Análisis de las fuentes: interpretación de los datos, valoración de su utilidad en el proyecto.
- Análisis de los campos, formatos y tipo de información de cada fuente y valoración del cruce de datos de distintas fuentes para nuestro proyecto.

1.1. Alcance (preliminar)

Explica brevemente qué información vamos a obtener de las distintas fuentes seleccionada y el uso que los datos podrían tener tras integrar y transformar las muestras con las que vamos a trabajar.

Identificador de la Ficha	**Búsqueda de fuentes**
Relatad todas las fuentes que habéis investigado y consultado en el proceso de localización de las que vais finalmente a utilizar.	*A la hora de buscar fuentes de datos como propuestas para nuestro proyecto, hemos tenido en cuenta como primera opción la base de datos de Valenbisi nos es proveída por el profesor de la UPV César Ferri, en la cual se nos ofrece información nueva de las estaciones de bicicletas en Valencia cada quince minutos (mediante un documento .csv). Así pues, para complementar la base de datos de Valenbisi, hemos decidido investigar acerca de las predicciones meteorológicas de las fechas que abarcan nuestros datos y por ello, hemos obtenido una base datos sobre la AEMET, que toman los datos desde la zona valenciana de Viveros, así como una base de datos que contiene información de zonas geográficas (en este caso barrios de Valencia con sus coordenadas) y otro dataset que estamos realizando para ubicar los lugares de interés de Valencia y relacionarlos con la estación Valenbisi más cercana teniendo en cuenta el cambio que este ofrece en al demanda de las bicis. Por último, queremos visualizar los cambios que se dan durante festivos en la demanda de bicicletas y por ello, hemos adquirido una base de datos acerca de los festivos.*

Criterios seguidos para la selección de las fuentes que se van a usar para el proyecto.	*Los criterios que hemos tenido en cuenta a la hora de seleccionar las fuentes de datos de nuestro proyecto han sido la capacidad de los datos totales para poder emplearlos al realizar los análisis que queremos de forma que podamos llegar, incluso, a agrupar los datos mediante clustering; la organización de las bases de datos en cuestión, de forma que entendamos todas sus variables y seamos capaces de obtener los análisis deseados; y que la base de datos contenga todo tipo de datos a la hora de interpretar sus valores al ponerlos en estudio, de forma que logremos obtener datos numéricos, categóricos, binarios...* *En cuanto a las bases de datos complementarias de nuestro dataset principal, hemos considerado que ayudarán a completar la información que ya teníamos y nos abra las puertas a nuevos análisis.*

1.2. Técnicas de obtención de datos y extracción

Explica las técnicas utilizadas en la obtención de las fuentes de datos. Especialmente, explica si has utilizado y para qué:

- Descarga de ficheros .csv
- Descarga desde una URL
- Lectura de tablas incrustadas en HTML
- Conversión de .JSON
- Conversión de .XML
- Recoger datos de Twitter y limpieza sobre expresiones regulares
- Recoger datos de Google y limpieza sobre expresiones regulares
- Web scraping
- Descargas en tiempo real durante varios ciclos

Para la obtención de los datos de Valenbisi, hemos hecho en primer lugar una descarga masiva de ficheros .csv (aproximadamente 8.000 archivos), los cuales corresponden a los datos entre el 1 de diciembre de 2022 y el 24 de febrero de 2023. Mediante varios programas (archivos .py de Python o scripts .sh del kernel de UNIX) hemos logrado tanto la unión como la separación de los datos dependiendo de las fechas y de las franjas horarias. Estos datos pueden descargarse desde la página de GitHub del profesor de la UPV César Ferri.

Con respecto a los datos meteorológicos, se ha creado un simple programa en Python que realiza solicitudes a la API de AEMET. Con la solicitud a dicha API se obtiene un primer diccionario (obtenido mediante la función json.loads) que contiene el link de los datos. Posteriormente, se realiza un proceso de scrapeo de los datos del link y se crea un segundo diccionario con los datos. Con esto, simplemente se crea un dataframe a partir del diccionario de datos y se exporta a csv.

Además, también hemos obtenido un fichero que delimita los barrios de Valencia mediante las coordenadas, lo que nos permitirá una mejor clasificación, agrupación y tratamiento de las estaciones de bicicletas que disponemos en el dataset.

De manera más rudimentaria, hemos creado varios ficheros auxiliares, como el de los festivos de Valencia (cuyos datos han sido obtenidos de …) o el fichero .xlsx que clasifica las estaciones según su importancia (hecho a mano usando el mapa que provee la compañía Valenbisi y un mapa normal de Valencia para determinar de manera manual si se encuentran cerca o no de un lugar de importancia).

1.3. Análisis de las fuentes: interpretación de los datos, valoración de su utilidad en el proyecto.

Por una parte, hemos descargado los ficheros .csv recogidos en el GitHub del profesor de la UPV César Ferri acerca de Valenbisi. Se trata de información que se actualiza con frecuencia, generando un fichero cada período corto de tiempo. De esta forma, lo que hemos hecho ha sido filtrar los datos desde el 1 de diciembre de 2022 hasta el 24 de febrero de 2023 para generar un único fichero .csv, que actúa como conjunto de datos de las estaciones de las bicis, aunque la idea es generar uno posterior separado por franjas horarias para facilitar y trabajar con análisis más específicos y con menos de cantidad de datos. Aquí encontramos información sobre la localización de las estaciones, así como el número de bicis disponibles o bornes libres, entre otros. Consideramos que esta base de datos tiene la máxima utilidad en nuestro proyecto porque es nuestra fuente de datos principal y, por ende, en la que se basa nuestro trabajo.

Por otra parte, hemos obtenido datos meteorológicos diarios de la estación meteorológica de Viveros en AEMET. Utilizando la API Key, nos devolvió una URL que contenía los datos y que metimos en un fichero .csv para poder trabajar mejor con ellos. Esta base de datos la utilizamos como complemento de la principal, de forma que podamos hacer una comparativa del uso de las bicicletas de Valenbisi cuando hay precipitaciones/mal tiempo y cuando no las hay. Así pues, este dataset tiene una gran utilidad por lo mencionado previamente.

Además, hemos encontrado un fichero que nos proporciona las coordenadas que delimita los barrios en Valencia, para ver las estaciones que se incluyen en cada barrio, y otro que nos permite visualizar los festivos en Valencia y como estos pueden afectar a la demanda de bicicletas. Estos dos datasets no son tan relevantes como los dos anteriores, pero también pueden ser muy útiles para el momento en el que quereamos realizar análisis diversos.

1.4. Análisis de los campos, formatos y tipo de información de cada fuente y valoración del cruce de datos de distintas fuentes para nuestro proyecto.

A la hora de interpretar la base de datos principal, podemos observar que tenemos las columnas, donde encontramos las variables: "*gid*", que representa la identificación de cada estación de bicis de Valenbisi; "*name*", que viene a ser la calle donde se sitúa la estación con un número de identificación que la acompaña; "*number_*", que representa el número de la estación y que podemos encontrar en la variable "*name*" también; "*address*", que sitúa la dirección de la estación; "*open*", que es una variable binaria que nos informa de si la estación está abierta o no; "*available*", que nos proporciona el número de las bicicletas disponibles en cada estación; "*free*", que es la variable complementaria de "*available*" y nos dice el número de bornes libres (o bicicletas en uso); "*total*", que nos informa del número total de bornes de cada estación; "*ticket*", que viene a ser una variable que contiene booleanos y nos dice si se puede pagar con tarjeta o no en dicha estación; "*updated_at*", que es la fecha de actualización del documento en cuestión a la base de datos; "*globalid*", que nos informa sobre la identificación "*global*" para cada estación; y "*geo_point_2*", que nos dice las coordenadas en cuanto a latitud y longitud de cada estación de Valenbisi. En cuanto a las filas, cada una representa una estación de Valenbisi, teniendo en la base de datos un total de 276 estaciones. Así pues, esta base de datos es el núcleo de nuestro proyecto.

En cuanto al .csv que hemos creado como base de datos de la AEMET, podemos diferenciar las siguientes variables entre las columnas: "*fecha*", "fecha del día (AAAA-MM-DD)"; "*indicativo*", "*indicativo climatológico*"; "*nombre*", "*nombre (ubicación)* de la estación"; "*provincia*", "*provincia de la estación*"; "*altitud*", "*altitud de la estación en m sobre el nivel del mar*"; "*tmed*", "*Temperatura media diaria*"; "*prec*", "*Precipitación diaria de 07 a 07*"; "*tmin*", "*Temperatura Mínima del día*"; "*horatmin*", "*Hora y minuto de la temperatura mínima*"; "*tmax*", "*Temperatura Máxima del día*"; "*horatmax*", "*Hora y minuto de la temperatura máxima*"; "*dir*", "*Dirección de la racha máxima*"; "*velmedia*", "*Velocidad media del viento*"; "*racha*", "*Racha máxima del viento*"; "*horaracha*", "*Hora y minuto de la racha máxima*"; "*sol*", "*Insolación*"; "*presmax*", "*Presión máxima al nivel de referencia de la estación*"; "*horapresmax*","*Hora de la presión máxima (redondeada a la hora entera más próxima)*"; "*presmin*", "*Presión mínima al nivel de referencia de la estación*"; "*horapresmin*", "*Hora de la presión mínima (redondeada a la hora entera más próxima)*".

Respecto al fichero de los barrios, nos encontramos con las siguientes variables: "*Codigo distrito-barrio*", el código del distrito/barrio; "*geo-shape*", las coordenadas que delimitan cada barrio; "*geo_point_2d*", punto central del barrio; "*Nombre*", nombre del barrio; "*Codigo barrio*", código del barrio; "*Codigo distrito*", código del distrito; "*Areas de barrios*", el área de cada barrio; "*LINKID*"; y, "*GLOBALID*". Además, hemos creado distintos ficheros con datos sobre las festividades en Valencia y sobre los lugares de interés que se encuentran cerca de las estaciones.

Las variables del primero tienen el siguiente significado: "*fecha*", fecha del día (DD/MM/AAAA); "*festivo*", si en esa fecha es festivo; "*vacaciones*", si hay vacaciones escolares; "*fin_de_semana*", si ese día es fin de semana; y, "*día_semana*", el día de la semana en el que se da esa fecha.

Referente al segundo, solo contiene dos variables: "*id*", el ID de cada estación; "*interés*", codificado del 0 al 8; y "metro" como una variable aparte a analizar, en el que se indica la existencia o no existencia de parada de metro o tranvía cercano. Los códigos significan lo siguiente:

0 – nada (no hay lugares de interés cercanos)
1 – colegio
2 – universidad
3 – instalación deportiva
4 – playa
5 – estación de trenes o autobuses
6 – centro comercial
7 – hospital
8 – otros