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:

Héctor	Beltrán Pozo
Juan José	Ruiz Esteban
Ivette	Mahmoud-Yousef Puig
Celia	Villar Arcos
Iván	Navarro Martínez

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.	**Calidad del aire de núcleos urbanos Europa –** Descarga de CSV *https://www.eea.europa.eu/en/topics/in-depth/air-pollution/european-city-air-quality-viewer* **Cantidad de coches por 1000 habitantes –** Web Scaping de tabla https://es.motor1.com/news/595127/lista-mercados-europeos-mas-coches/ **Calidad del aire en radio aprox. –** API request a GoogleApis con geolocator https://developers.google.com/maps/documentation/air-quality?hl=es_419 **Cantidad de población –** Descarga de CSV https://ec.europa.eu/eurostat/databrowser/view/tps00001/default/table?lang=en&category=t_reg.t_reg_dem **Cantidad de aeropuertos en ciudades europeas –** Descarga de CSV https://www.uv.es/dataenhance/datasets/000017.html **Cantidad de transporte aéreo (partidas de vuelos en todo el mundo) –** Web Scraping de tabla https://datos.bancomundial.org/indicador/IS.AIR.DPRT?most_recent_year_desc=true

Criterios seguidos para la selección de las fuentes que se van a usar para el proyecto.	Las fuentes seleccionadas para el análisis de la calidad del aire en ciudades europeas han sido elegidas por su fiabilidad, actualización y cobertura geográfica. Se emplean organismos oficiales como la EEA, Eurostat y el Banco Mundial, que ofrecen datos revisados y estandarizados. La API de Google proporciona información precisa a nivel de coordenadas, mientras que fuentes con datos estructurados en CSV facilitan su procesamiento. En casos de datos no disponibles en formatos abiertos, se emplea web scraping, asegurando su correcta extracción y limpieza. La combinación de estas fuentes permite un análisis detallado y riguroso.

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:

- **Web Scraping:** extracción de datos de páginas web sin API ni CSV disponibles, utilizando parsing de HTML (BeautifulSoup). Aplicado a datos de coches por 1000 habitantes y vuelos. Para extraer datos publicados por revistas que no tienen acceso directo, pero que son relevantes para el análisis.
- **API Request:** obtención automatizada de datos en tiempo real mediante solicitudes a una API. Usado en Google APIs Air Quality con geolocalización para calidad del aire. Para la obtención de datos en un tiempo que nosotros podamos determinar, para hacer que los datos coincidan entre sí a nivel temporal.
- **Descarga de CSV:** acceso a datos estructurados desde fuentes oficiales como EEA, Eurostat y UV. Facilita procesamiento y análisis sin necesidad de extracción adicional. Es la forma más eficiente de datos que ya han sido trabajados y estandarizados.

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

Cada fuente de datos ha sido analizada en función de su relevancia para el estudio de la calidad del aire en ciudades europeas, considerando su utilidad, precisión y aplicabilidad en el proyecto:

- **Calidad del aire (EEA, Google APIs)**
  - *Interpretación:* proporciona valores de contaminantes clave (NO_2, PM2.5, O_3) en diferentes ubicaciones.
  - *Utilidad:* esencial para evaluar los niveles de contaminación y correlacionarlos con factores urbanos y geográficos.
- **Cantidad de coches por 1000 habitantes (Web Scraping de Motor1.com)**
  - *Interpretación:* indica la densidad de vehículos privados, un factor relevante en la emisión de contaminantes.
  - *Utilidad:* clave para analizar el impacto del tráfico en la calidad del aire.
- **Población (Eurostat)**
  - *Interpretación:* proporciona datos demográficos de ciudades europeas.
  - *Utilidad:* necesario para normalizar datos de contaminación y transporte en función del tamaño poblacional.
- **Aeropuertos y tráfico aéreo (UV, Banco Mundial)**
  - *Interpretación:* permite conocer la cantidad de infraestructuras aeroportuarias y el volumen de vuelos en cada ciudad.
  - *Utilidad:* relevante para evaluar el impacto del transporte aéreo en la contaminación urbana.

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. Valoración del cruce de datos de distintas fuentes para nuestro proyecto.

Para garantizar la coherencia y utilidad del análisis, se han evaluado los formatos, tipos de información y la posibilidad de combinar datos de distintas fuentes.

Campos y formatos de cada fuente

- **Calidad del aire (EEA, Google APIs)**
  - *Campos:* niveles de NO_2, PM2.5, O_3; ubicación geográfica (latitud/longitud); fecha y hora de medición.
  - *Formato:* CSV (EEA), JSON (Google APIs).
  - *Tipo:* datos numéricos y geoespaciales.
- **Cantidad de coches por 1000 habitantes (Motor1.com – Web Scraping)**
  - *Campos:* país, número de coches por cada 1000 habitantes.
  - *Formato:* tabla extraída en CSV o DataFrame.
  - *Tipo:* datos numéricos y categóricos.
- **Población (Eurostat – CSV)**
  - *Campos:* país, ciudad, número de habitantes.
  - *Formato:* CSV.
  - *Tipo:* datos numéricos.
- **Aeropuertos y tráfico aéreo (UV, Banco Mundial – CSV, Web Scraping)**
  - *Campos:* ciudad, número de aeropuertos, volumen de vuelos por año.
  - *Formato:* CSV (aeropuertos), tabla web extraída como CSV (vuelos).
  - *Tipo:* datos numéricos.

Valoración del cruce de datos

El cruce de datos entre distintas fuentes permite enriquecer el análisis de la calidad del aire:

- Relacionar niveles de contaminación (EEA, Google APIs) con densidad de vehículos (Motor1.com) y población (Eurostat) para evaluar el impacto del tráfico en la polución urbana.
- Analizar el efecto del transporte aéreo cruzando datos de aeropuertos y vuelos con contaminación atmosférica en ciudades cercanas.
- Aprovechar coordenadas geográficas de calidad del aire para mapear información de otras fuentes (densidad de población, infraestructuras de transporte).