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:

Nadal	Bardisa Quintero
María	Martínez Fernández
Antonio	Sánchez Guillén
Sergio	Vidal Picazo
Isabelle	Archer

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.	*Se va a realizar una búsqueda abierta a todo tipo de proveedores de datos, como el Banco Mundial, el Instituto Nacional de Estadística, el Fondo Monetario Internacional, etcétera.* *Tras un estudio de los datos ofrecidos por estas fuentes, se han escogido los proporcionados por Inside Airbnb, centrando el estudio y análisis a los datos obtenidos, y tomando datos añadidos de Idealista para completar el análisis con comparaciones entre los precios de Airbnb y de viviendas en alquiler o venta.* *Por comodidad y tiempo se ha puesto foco únicamente en viviendas completas, ignorando los alquileres de habitación. Además, se va a realizar el análisis inicialmente sólo de Valencia, planteando la posible extensión a otras ciudades con el servicio de alquiler de Airbnb.*

Criterios seguidos para la selección de las fuentes que se van a usar para el proyecto.	*Se han buscado fuentes de datos que proporcionen archivos de formato csv o xlsx para facilitar la lectura de estos.* *Se ha centrado la búsqueda en datos proporcionados por la página web de Inside Airbnb para obtener información de las viviendas y sus reseñas.* *Se ha dado importancia a las fuentes que tienen bastantes datos y provienen de compañías conocidas y reputadas. Por ejemplo, Airbnb es muy conocida y tiene datos de muchos años. También Idealista la usa mucha gente y lleva un gran registro de datos sobre las viviendas.* *Estos datos van a permitir encontrar relaciones entre variables que permitan a los usuarios que buscan vivienda elegir una que se ajuste a su presupuesto y supla sus necesidades.* *Mediante un análisis en profundidad de los comentarios (análisis de sentimiento) se puede obtener información valiosa sobre las valoraciones en la plataforma de dicha vivienda, permitiendo predecir indistintamente a partir del resto de variables el precio, la positividad de los comentarios y la valoración (1-5) de la vivienda.*

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

Se han obtenido los datos de la descarga de dos ficheros .csv (listings.csv y reviews.csv), para cada ciudad (inicialmente de Valencia), de la página de Inside Airbnb, que proporciona los datos suficientes para llevar a cabo un análisis extenso en cuanto al efecto de diversas variables en el precio, el sentimiento de las reseñas, e incluso detectar el efecto de variables externas como los cambios de leyes o políticas de alquiler de vivienda.

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

Para cada ciudad que vayamos a analizar tenemos un fichero para los Airbnbs de esa ciudad y otro que incluye sus reseñas. El fichero de Airbnb primero se ha filtrado por la variable property_type, ya que el Airbnb podía ser un piso o solo una habitación y el precio de una a otra difiere. Finalmente se han seleccionado solo los que son pisos completos. Este fichero contiene información sobre las características del piso (habitaciones, baños ….), además de características del host, estadísticas de las reservas, etc.

El fichero de reseñas complementa con información sobre los clientes que han puesto la reseña, la fecha en la que fue publicada y el comentario que hacen respecto a ella.

Cruzando ambos datasets, se pueden identificar patrones entre la calidad de los alojamientos y sus precios, o evaluar cómo influyen las reseñas en la popularidad de un alojamiento.

Permiten realizar análisis descriptivos (promedio de precios, puntuaciones, distribución por barrios) y análisis más avanzados como modelos predictivos o análisis de sentimientos en las reseñas.

Se pueden encontrar algunas limitaciones como posibles datos faltantes, sesgos en las reseñas (opiniones extremas), y la necesidad de limpieza y normalización de los datos antes del análisis.

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.

Para cada ciudad, la página Inside Airbnb proporciona varios ficheros .csv.gz. Hay un csv para las viviendas, otro para el calendario, y el último para las reseñas. Se puede combinar el fichero de las viviendas y de las reseñas concatenando los datos con la columna "id" en el listings.csv y con la columna "listing_id" en el reviews.csv. Así que, se añaden las reseñas a las viviendas que las tengan.

Esto servirá para realizar algunos estudios sobre la vivienda, como análisis de sentimiento a partir de las reseñas, comparación de precios y oferta según la zona o comprobar si ha cambiado la cantidad de alquileres según los cambios de política.

Listings.csv:

Es un fichero .csv con 75 variables y 8699 observaciones. Después de mirar todas las variables, se han quitado varias variables por distintos motivos: incluía un url que no vamos a analizar, era una variable derivada y presentaba correlación que podría afectar al futuro modelo, o la variable no varía. La tabla siguiente (*tabla 1*) describe las variables con las que quedamos para las viviendas:

Variable	Descripción	Tipo	Valores faltantes
id	Identificador de la vivienda	texto	No
name	Nombre de la vivienda	texto	No
description	Descripción de la vivienda	texto	No
neighborhood_overview	Breve descripción de la situación del barrio	texto	Sí
host_id	Identificador del propietario de la vivienda	texto	No
host_name	Nombre del propietario de la vivienda	texto	No
host_since	Fecha de registro del propietario	fecha/texto	No
host_location	Ubicación geográfica del propietario	texto	Sí
host_about	Texto de presentación del propietario en su perfil	texto	Sí
host_response_time	Tiempo de respuesta del propietario al arrendatario	texto	Sí
host_response_rate	Porcentaje de preguntas de arrendatarios respuestas por el propietario	numérica	Sí
host_acceptance_rate	Porcentaje de aceptación de arrendatarios por parte del propietario	numérica	Sí
host_is_superhost	El propietario es valorado muy positivamente	booleano	Sí
host_neighbourhood	Nombre del barrio en el que reside el propietario	texto	Sí
host_listings_count	Cantidad de viviendas ofertadas por el propietario actualmente	numérica	No

host_total_listings_count	Número de viviendas distintas ofertadas por el propietario históricamente	numérica	No
host_verification	Lista de los distintos métodos de verificación de identidad del propietario	texto	No
host_has_profile_pic	El propietario presenta una foto de perfil	booleano	No
host_identity_verified	El propietario tiene su identidad verificada	booleano	No
neighbourhood	Municipio de la vivienda	texto	Sí
neighbourhood_cleansed	Barrio de la vivienda	texto	No
neighbourhood_group_cleansed	Distrito de la vivienda	texto	No
latitude, longitude	Coordenadas de la vivienda	numérica	No
accommodates	Número de personas que la propiedad puede alojar cómodamente	numérica	No
bathrooms	Número de baños en la vivienda	numérica	Sí
bedrooms	Número de habitaciones en la vivienda	numérica	Sí
amenities	Lista de servicios y comodidades disponibles en la propiedad	texto	Sí
price	Precio de la vivienda	numérica	Sí
minimum_nights	Mínimo número de noches a contratar	numérica	No
maximum_nights	Máximo número de noches a contratar	numérica	No
availability_365	Días al año que se encuentra disponible ese espacio	numérica	No
number_of_reviews	Número de reseñas	numérica	No
number_of_reviews_ltm	Número de reseñas en los últimos 12 meses	numérica	No
first_review	Fecha de la primera reseña	texto/fecha	Sí
last_review	Fecha de la última reseña	texto/fecha	Sí
review_scores_rating	Valoración del rating atribuido a la vivienda previa-estancia de los arrendatarios	numérica	Sí
review_scores_accuracy	Valoración del acertamiento de los comentarios respecto de la vivienda	numérica	Sí

review_scores_cleanliness	Valoración de la limpieza de la vivienda por parte de los arrendatarios	numérica	Sí
review_scores_checkin	Valoración del recibimiento de los arrendatarios por parte del propietario	numérica	Sí
review_scores_communication	Valoración de la comunicación entre el propietario y los arrendatarios	numérica	Sí
review_scores_location	Valoración de la ubicación de la vivienda por parte de los arrendatarios	numérica	Sí
review_scores_value	Valoración de la relación calidad-precio por parte de los arrendatarios	numérica	Sí
license	Licencia identificativa del propietario	texto	Sí
instant_bookable	La propiedad puede reservarse de forma inmediata	booleano	No
Reviews_per_month	Número de reseñas por mes	numérica	Sí

Tabla 1: descripción de las variables de listings.csv.

Reviews.csv:

Variable	Descripción	Tipo	Valores faltantes
listing_id	Identificador de la reseña generado dentro de la tabla	categórica	No
id	Número de identificación de la reseña	categórica	No
date	Fecha de publicación de la reseña	texto/fecha	No
reviewer_id	Número de identificación del cliente	texto	No
reviewer_name	Nombre del cliente	texto	No
comments	La reseña publicada por el cliente	texto	Sí

Tabla 2: descripción de las variables de reviews.csv.