Universidad de los Andes > FACULTAD DE INGENIERÍA Y CIENCIAS APLICADAS

Entrega N°1

Almacenamiento y Procesamiento Masivo de Datos

"Análisis de las viviendas e historial de arriendos en AirBnb"

Alumnos:
Andrés Abbott
Juan Pablo Arévalo

Profesor: Ignacio Pérez

10/09/2017

Contexto

Airbnb fue fundada el año 2008 por Brian Chesky, Joe Gebba y Nathan Blecharczyk y es actualmente la plataforma líder en arriendo de viviendas y habitaciones a lo largo del mundo.

Elegimos trabajar con el set de datos de AirBnb, plataforma que presenta una oferta de alrededor de 2.000.000 de propiedades activas en 192 países y 33.000 ciudades. A la fecha se han registrado alrededor de 100.000.000 de reservas a través de la plataforma, lo cual es constituido por miles de reservas cada mes.

Índice

Recolección de datos

Los datos de AirBnb pueden descargarse por medio de un Scraper que extrae los datos directamente desde la página web.
Existen actualmente datasets previamente descargados vía Scraping que son abiertos al público para descargarlos[1]. Por motivos de facilidad, utilizamos estos datasets.
Para tener una gran cantidad de datos y establecer comparaciones interesantes entre países, elegimos descargar los datasets para 5 ciudades distintas:

- Amsterdam, Países Bajos
- Dublín, Irlanda
- Hong Kong, China
- New York, EEUU
- París, Francia

Los datos se encuentran todos actualizados al año 2017, a excepción de los datos de Hong Kong que se actualizaron en Agosto del año 2016.

[1] http://insideairbnb.com/get-the-data.html

Estructura de datos

A continuación se procederá a explicar la estructura de las tablas con las que trabajaremos

- Calendario: Contiene información general de los listings por fecha y disponibilidad de arriendo.
    - Estructura : Consta de 4 columnas
        - Listing_id
        - Date
        - Available : disponibilidad de listing. Puede ser f para falso y t para true
        - Price: precio el listing

- Reviews: Contiene la información de los reviews para cada listing. Dicha informacion se complementará de mejora manera con los datos dentro de la tabla listing. Cabe destacar que en esta tabla no se guardan las puntuaciones, solo los comentarios de los reviews.
    - Estructura: Consta de 6 columnas
        - Listing_id
        - Id
        - Date
        - reviewer_id
        - reviewer_name
        - comments

- Listing: Tabla principal con la que trabajaremos. Contiene la información principal de cada listing.
    - Estructura: Consta con más de 40 columnas
        - Id
        - Listing_url
        - Scrape_id : usuario que descargó los datos de Airbnb
        - Last_scraped: fecha en que se descargaron los datos de Airbnb
        - Name
        - Summary
        - Space: descripción del espacio de la vivienda.
        - Description
        - Experiences_offered : exprerienciias extra que ofrece el listing, como ayuda con el idioma.
        - Neighborhood_overview: descripción del vecindario.
        - Notes
        - Transit: tránsito y locomoción.

- Access: acceso al espacio arrendado.
- Interaction: interacción con los propietarios.
- House_rules: reglas de la casa
- Thumbnail_url
- Medium_url
- Picture_url
- Xl_picture_url
- Host_id
- Host_url
- Host_name: nombre del propietarios
- Host_since
- Host_location
- Host_about: descripción del propietario
- Host_response_time: tiepo de respuesta del propietario
- Host_response_rate: porcentaje de respuesta del propietario
- Host_is_superhost: se desconoce el significado de esta columna.
- Host_thumbnail_url
- Host_picture_url
- Host_neighbourhood
- Host_listings_count
- Host_total_listing_count : cantidad total de arriendos.
- Host_verification : Verificaciones, pueden ser teléfonos, email, facebook, etc.
- Host_has_profile_pic
- Host_identify_verified
- Street
- Neighbourhood
- Neighbourhood_cleansed
- Neighbourhood_group_cleansed
- City
- State
- Zipcode
- Market
- Smart_location
- Country_code
- Country
- Latitude
- Longitude
- Is_location_exact
- Property_type: si es departamento, casa, etc.
- Room_type: tipo de pieza a utilizar.
- Accommodates: cantidad de personas que puede albergar.
- Bathrooms
- Bedrooms
- Beds

- Bed_type
- Amenities: implementos que trae, como tv, cocina, etc.
- Square_feet: metros cuadrados del espacio.
- Price: precio
- Weekly_price
- Monthly_price
- Security_deposit: depósito de seguridad por adelantado.
- Cleaning_fee
- Guests_included
- Extra_people
- Minimum_nights
- Maximum_nights
- Calendar_updates
- Has_availability
- Availability_30
- Availability_60
- Availability_90
- Availability_360
- Calendar_lat_scraped
- Number_of_reviews
- First_review
- Last_review
- Review_score_rating
- Review_score_accuracy
- Review_score_cleanlinnes
- Review_score_checkin
- Review_score_communication
- Review_score_location
- Review_score_value
- Requieres_license: se desconoce el significado de esta columna.
- License: se desconoce el significado de esta columna.
- Juridiction_name
- Cancelation_policy: política de cancelación.
- Review_per_month

Propuesta de Consultas

A continuación presentamos una propuesta con las consultas y métricas que podrían resultar interesantes a la hora de analizar los datos obtenidos de AirBnb. Separamos dichas métricas en dos grupos: Métricas que se calculan para cada ciudad y Métricas generales.

Métricas que se calculan para cada ciudad:

1) Listing con mayor precio promedio.
2) Listing con menor precio promedio.
3) Listing con mayor número de reviews.
4) Listing que se ha arrendado más veces.
5) Listings disponibles entre un rango de fechas.
6) Listing más barato disponible entre un rango de fechas.
7) Listing con mejor rating de los listings más baratos:
    - El objetivo es encontrar el listing mejor evaluado de los más baratos.
8) Listing más barato entre los listings con mejor rating:
    - El objetivo es encontrar el listing más barato de los mejores evaluados.
9) Listing con reviews más negativos:
    - Se buscarán palabras en los comentarios como “unhappy”, “unpleasant”, “worst”, “disappointed”, etc.
10) Usuario (host) con más listings.
11) Usuario con mayor número de reviews.
12) Usuario más disconforme en sus reviews:
    - Se buscarán palabras en los comentarios como “unhappy”, “unpleasant”, “worst”, “disappointed”, etc.
13) Usuarios que se arrienden listings mutuamente:
    - Sería interesante identificar si un par de usuarios se arrendaron una vivienda mutuamente, y de ser así, encontrar el par con mayor número de arriendos mutuos.
14) Par de Usuarios con mayor similitud en listings visitados:
    - Se busca encontrar usuarios con similitudes en sus arriendos, para luego poder recomendar listings en base a los usuarios comunes.
15) Arriendo que generó más ingresos para su host:
    - Esto se encontraría calculando el valor del arriendo multiplicado por el número de días que se arrendó.
16) Porcentaje de arriendos en fechas festivas:
    - Se busca comparar el total de arriendos con el número de arriendos para determinadas fechas festivas, por ejemplo año nuevo, navidad, etc.

17) Promedio del valor de cada cama:
    - Esto se encontraría calculando un promedio sobre todos los datos, considerando el valor de la vivienda (listing) comparado con el número de camas presentes en esta.
18) Promedio del número de baños en relación a los huéspedes:
    - Se encontraría calculando un promedio sobre todos los datos, considerando el número de baños y de huéspedes en cada vivienda.
19) Número de listings por barrio.

Métricas generales:

1) Usuario con listings en más países:
    - El objetivo es encontrar el usuario que presente listings en más países del set descargado.
2) Usuario con reviews en más países:
    - Similar al anterior, se busca encontrar al usuario que presente reviews en más países, por ende, el usuario que ocupó un listing en más países.
3) Comparativa de precio promedio entre países:
    - Realizar una comparación entre el precio promedio de listings para los países del set de datos, concentrándose en viviendas de características similares y poder así llegar a conclusiones como “una cama en Nueva York equivale a 5 camas en Dublín”, por ejemplo.
    -

Dificultades en el análisis de los datos

Luego de trabajar durante un tiempo con los datos, nos topamos con algunas dificultades que podrían perjudicar el proceso de cálculo de las métricas previamente mencionadas. Explicaremos dichas dificultades a continuación:

1) Presencia del carácter ‘,’ (coma) en algunos campos:
    - Como MRJob entrega línea por línea el contenido del archivo .csv, al intentar separar por comas, se consideran substrings de los comentarios de reviews o descripciones de listings como campos de la tupla, ya que el comentario presentaba comas dentro de su contenido. Esto podría solucionarse si en lugar de utilizar MRJob se abriera el .csv directamente y se separan los campos basándose en la cabecera. Puede que MRJob también entregue una solución a este problema, y eso es lo que investigaremos en las siguientes entregas.

2) Presencia de saltos de línea en algunos campos:
    - El problema es bastante similar al anterior, los comentarios de reviews o descripciones de los listings generalmente presentan saltos de línea (‘\n’). MRJob interpreta estos saltos de línea como una nueva tupla, por lo que se generan tuplas falsas con contenidos que no se asocian al header original. La solución es análoga a la anteriormente mencionada, es decir, abrir el archivo .csv manualmente y obtener las líneas separadas ignorando los saltos de línea dentro de los campos o si no estudiar más sobre el funcionamiento de MRJob y comprobar si existe una forma de manejar este problema.

3) Poca claridad o documentación sobre la estructura de los datos.
    - Existen muchos campos con nombres ambiguos y no documentados (o al menos, a la fecha no encontrados por nosotros) que dificultan el entendimiento del contenido de estos. Por ejemplo los campos “market” o “neighborhood_host” de listings que no presentan suficiente información como para inferir a que se refieren. Esto se podría solucionar investigando a mayor profundidad la estructura de datos de AirBnb y relacionándola con los datos descargados.