

Trabajo Práctico 1: Análisis Exploratorio de Datos

[75.06 / 95.58] Organización de Datos
Segundo cuatrimestre de 2018

Grupo Datatouille

Alumno	Padrón	Mail
del Mazo, Federico	100029	delmazofederico@gmail.com
Bojman, Camila	101055	camiboj@gmail.com
Hortas, Cecilia	100687	ceci.hortas@gmail.com
Souto, Rodrigo	97649	rnsoutob@gmail.com

<https://github.com/FdelMazo/7506-Datos/>
<https://kaggle.com/datatouille2018/7506-TP1/>

Curso 01

- Argerich, Luis Argerich
- Golmar, Natalia
- Martinelli, Damina Ariel
- Ramos Mejia, Martín Gabriel

Contents

1	Introducción	1
2	Información general sobre los datos	1
3	Datasets adicionales incorporados para el análisis	2
4	Ejecución	2
5	Análisis de eventos	3
5.1	Hipótesis sobre el truncamiento de los datos	3
5.2	Frecuencia de eventos	4
6	Conclusiones, insights y aportes	4

1 Introducción

Se propone analizar en el presente informe los datos obtenidos de usuarios que visitaron `www.trocafone.com`, un sitio de e-commerce de compra y venta de celulares reacondicionados, con operaciones principalmente en Brasil. Para ello, la empresa Trocafone nos proporcionó acceso a los datos a través del archivo `events.csv`.

El objetivo de este informe es realizar un análisis de dicho set de datos para obtener un listado de *insights* aprendidos sobre los mismos. Se propone específicamente:

- Descubrir features en el campo `model`
- Identificar patrones de usuarios que realizan checkouts y conversiones
- Analizar las búsquedas que realizan los usuarios y las *keywords* utilizadas
- Analizar los distintos lugares de dónde se originan las visitas a Trocafone
- Descubrir features jerarquizando alguno de los campos disponibles
- **agregar más items a medida que desarrollamos el informe**

Finalmente se busca realizar un aporte a la empresa Trocafone con datos que sirvan para mejorar sus servicios.

2 Información general sobre los datos

En el archivo proporcionado por la empresa Trocafone se observan las siguientes columnas:

- **timestamp**: Fecha y hora cuando ocurrió el evento.
- **event**: Tipo de evento.
- **person**: Identificador de cliente que realizó el evento.
- **url**: Url visitada por el usuario.
- **sku**: Identificador de producto relacionado al evento.
- **model**: Nombre descriptivo del producto incluyendo marca y modelo.
- **condition**: Condición de venta del producto.
- **storage**: Cantidad de almacenamiento del producto.
- **color**: Color del producto.
- **skus**: Identificadores de productos visualizados en el evento.
- **search_term**: Términos de búsqueda utilizados en el evento.
- **static_page**: Identificador de página estática visitada.

- **campaign_source**: Origen de campaña, si el tráfico se originó de una campaña de marketing.
- **text_engine**: Motor de búsqueda desde donde se originó el evento, si aplica.
- **channel**: Tipo de canal desde donde se originó el evento.
- **new_vs_returning**: Indicador de si el evento fue generado por un usuario nuevo (New) o por un usuario que previamente había visitado el sitio (Returning) según el motor de analytics.
- **city**: Ciudad desde donde se originó el evento.
- **region**: Región desde donde se originó el evento.
- **country**: País desde donde se originó el evento.
- **device_type**: Tipo de dispositivo desde donde se generó el evento.
- **screen_resolution**: Resolución de pantalla que se está utilizando en el dispositivo desde donde se generó el evento.
- **operating_system_version**: Versión de sistema operativo desde donde se originó el evento.
- **browser_version**: Versión del browser utilizado en el evento.

3 Datasets adicionales incorporados para el análisis

Se utilizan adicionalmente los siguientes archivos para realizar el análisis:

- Mapas de Brazil y Estados Unidos: Sacados de Geonames ¹
- `brands.csv` para extraer información del modelo ²
- `os.csv` para extraer información del sistema operativo ³
- `browsers.csv` para extraer información de la versión del browser ⁴

4 Ejecución

El trabajo fue realizado en Anaconda ⁵. Para poder replicar el trabajo, hay que también instalar las siguientes librerías adicionales:

- Squarify ⁶: Para los treemaps.
- Geopandas ⁷: Para poder graficar sobre mapas geográficos.

¹<http://www.geonames.org/>

²Detallado en `anexo.ipynb`

³Detallado en `anexo.ipynb`

⁴Detallado en `anexo.ipynb`

⁵<https://anaconda.org/>

⁶<https://github.com/laserson/squarify>

⁷<http://geopandas.org/>

- Wordcloud ⁸: Para poder visualizar los términos más buscados.

Estos pueden ser instalados con los siguientes comandos:

```
pip install squarify
conda install -c conda-forge geopandas
conda install -c conda-forge wordcloud
```

5 Análisis de eventos

En esta sección se propone analizar los distintos tipos de eventos realizados por los usuarios de Trocafone.

El campo **event** puede adquirir distintos tipos de valores categóricos que se describen como sigue:

- **"viewed product"**: El usuario visita una página de producto.
- **"brand listing"**: El usuario visita un listado específico de una marca viendo un conjunto de productos.
- **"visited site"**: El usuario ingresa al sitio a una determinada url.
- **"ad campaign hit"**: El usuario ingresa al sitio mediante una campana de marketing online.
- **"generic listing"**: El usuario visita la homepage.
- **"searched products"**: El usuario realiza una búsqueda de productos en la interfaz de búsqueda del site.
- **"search engine hit"**: El usuario ingresa al sitio mediante un motor de búsqueda web.
- **"checkout"**: El usuario ingresa al checkout de compra de un producto.
- **"static page"**: El usuario visita una página.
- **"conversion"**: El usuario realiza una conversión, comprando un producto.
- **"lead"**: El usuario se registra para recibir una notificación de disponibilidad de stock, para un producto que no se encontraba disponible en ese momento.

5.1 Hipótesis sobre el truncamiento de los datos

Se sabe por información de la consigna que el dataset proporcionado no representa el conjunto total de datos de todos los eventos realizados por los usuarios en el período de tiempo determinado. Es por esta razón que se busca elaborar una hipótesis en base al criterio con el que se fijó la selección de los datos. A partir de un análisis de los mismos se observó que todos los usuarios registrados en el dataset realizaron al menos 1 checkout, por lo que la base de datos original

⁸https://github.com/amueller/word_cloud/

se truncó. Sin ir más lejos es evidente que no todos los usuarios que ingresan al sitio de Trocafone van a realizar un checkout.

De esta manera es importante remarcar que las conclusiones a las que se llegará en el desarrollo del Trabajo Práctico se basan en un sector segmentado de los datos, por lo que en ciertos aspectos del análisis no se podrá arribar a conclusiones fundadas.

5.2 Frecuencia de eventos

Se analiza qué tipo de evento es el más frecuente en el dataset. Para ello se grafica la cantidad registrada de eventos en función de los distintos tipos de eventos. Se excluyen del análisis los eventos **checkout** y **conversion** ya que serán analizados en detalle a lo largo del informe y los eventos **lead** y **static page** ya que no se consideran relevantes al análisis que se quiere realizar.

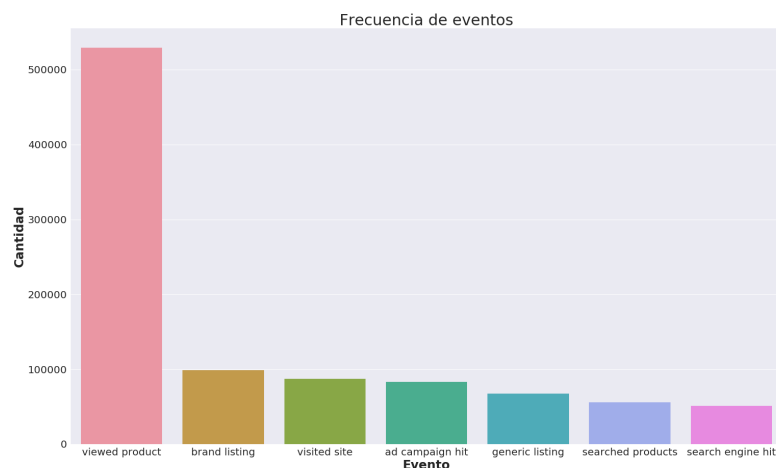


Figure 1: Frecuencia de eventos

Se observa en el gráfico que la mayor cantidad de eventos se relacionan a la vista de un producto, lo cual era previsible ya que Trocafone es una plataforma de e-commerce y ver productos es su principal función como sitio.

6 Conclusiones, insights y aportes

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Proin ultricies justo nisi, in ultrices lorem sollicitudin sed. Donec diam velit, aliquet et neque ac, tempus condimentum dolor. Maecenas scelerisque malesuada dignissim. Morbi sollicitudin est eu varius vestibulum. Duis sollicitudin non sapien quis iaculis. Quisque et luctus massa. In vitae odio vitae erat dapibus laoreet vitae in massa. Interdum et malesuada fames ac ante ipsum primis in faucibus. Nunc eu nunc tellus. Proin dignissim venenatis justo, vel rhoncus nibh commodo tincidunt.

Suspendisse eleifend massa eget ligula viverra lacinia. Mauris egestas nisl a tincidunt rhoncus.