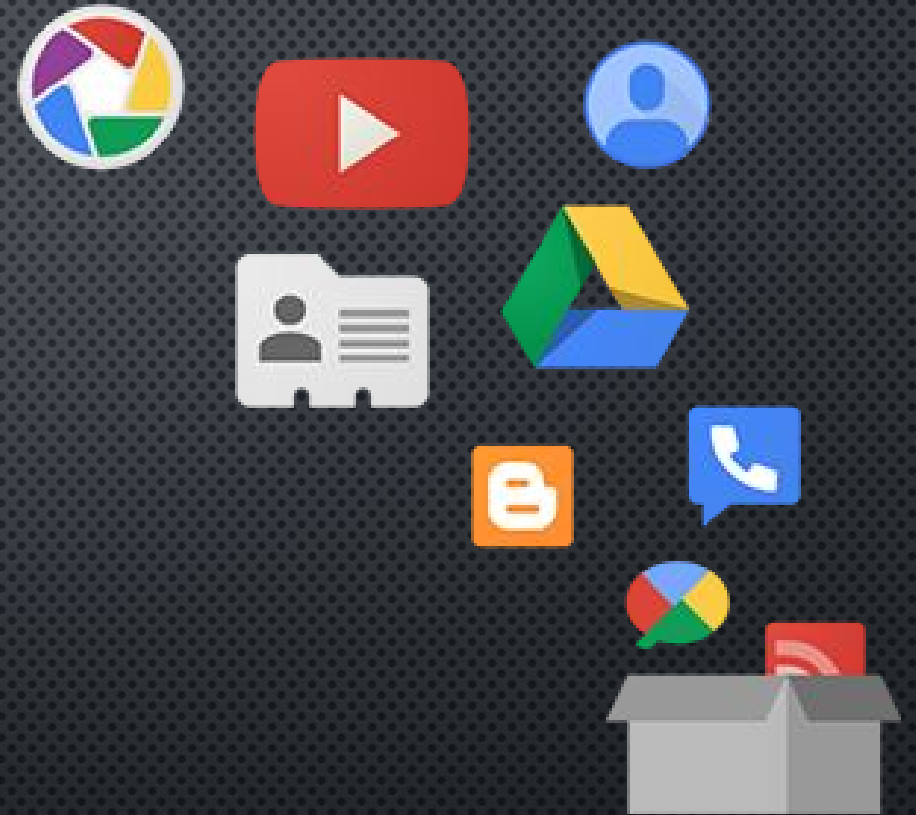


GOOGLE TAKEOUT



Alfredo, Alaín, Eduardo, Pedro V

Un poco de historia...









Google Takeout es un proyecto impulsado por un grupo de ingenieros progresistas de Google (Google Data Liberation Front) que pretende hacer de manera fácil que cualquier usuario pueda exportar los datos de sus aplicaciones.

La primera versión fue en 2011 liberando buzz, circles y profile, eventualmente las demás se fueron agregando, como Gmail en 2013 y así sucesivamente.

Actualmente se pueden descargar datos de 23 aplicaciones de Google, cada uno con un formato específico.

Se trata de la información generada de manera personal, para cada cuenta de Google(*) a través de los servicios de:

-  Búsquedas de google 
-  Correos electrónicos 
-  Historial de ubicaciones 

- La cantidad de datos varían para cada cuenta.
- Por los casos explorados, se cuenta con datos desde 2011.
- El campo en común es el timestamp.
- Nos ayudará a generar relaciones entre apps

Búsquedas

```
{"query":  
  {"id":  
    [{"timestamp_usec":"1407774749032392"}],  
    "query_text":"banco mundial"}}  
{"query":  
  {"id":  
    [{"timestamp_usec":"1407774749075527"}],  
    "query_text":"data lake"}}  
{"query":  
  {"id":  
    [{"timestamp_usec":"1407774749095273"}],  
    "query_text":"shiba dog"}}
```

¿Cómo son las búsquedas por hora,
día, mes?
¿Existen tiempos prolongados de
búsqueda?
¿Búsquedas productivas?
¿Dicen algo las palabras más
buscadas?



Ubicaciones

- ¿Cual es la frecuencia de movimientos?
- ¿Se puede identificar trabajo y hogar?
- ¿Cuando se identifica una mudanza?
- ¿Cuales son los traslados promedio en tiempo y distancia?

```
{
  "timestampMs": "1414819151315",
  "latitudeE7": 204435729,
  "longitudeE7": -872882348,
  "accuracy": 49,
  "activitys": [ { "timestampMs": "1414819136573",
    "activities": [ { "type": "inVehicle", "confidence": 62 },
      { "type": "still", "confidence": 29 },
      { "type": "onBicycle", "confidence": 5 },
      { "type": "unknown", "confidence": 5 }
    ]
  } ]
}
```



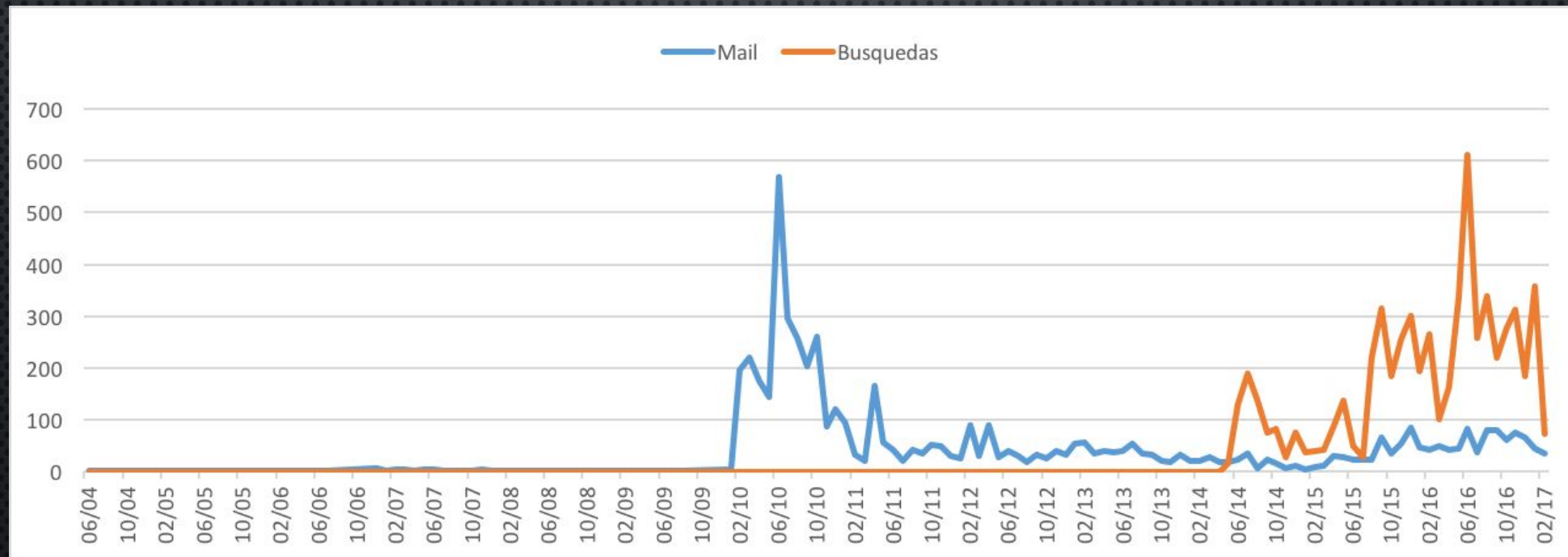
Correos

X-GM-THRID: 1545043292255087830
X-Gmail-Labels: Importante, Destacados, Recibidos
From: <ventasweb@interjet.com.mx>
To: <xxx@gmail.com>
Reply-To: <ventasweb@interjet.com.mx>
Date: Fri, 9 Sep 2016 19:41:43 -0500
Subject: Interjet Itinerario
Content-Type: multipart/alternative;
Message-ID: <e34b7917-c506-4a84-ac90-626bf8fafb7a
Content-Transfer-Encoding: quoted-printable
—CONTENT—

- ¿Cual es el tráfico a través del tiempo?
- ¿Es posible hacer una red de personas?
- ¿Cual es la relación de enviados con recibidos?
- ¿Alguna relación de los dominios con SPAM?
- ¿Dice algo especial el asunto?



Exploración



Producto: Sistema de Recomendación

- Estimar algún nivel de zona (ciudad, C.P., entidad, país, etc) en el que el usuario haya vivido o trabajado.
- Identificar, a través de ubicaciones, gustos o actividades adicionales del usuario.
- Aproximar el perfil del usuario vinculando sus localizaciones con sus búsquedas en Google y/o información contenida en sus correos electrónicos.
 - Edad, Género, Actividad laboral
- Estimar a qué se dedica el usuario, mediante la correlación de correos y búsquedas.
- Para esto necesitamos una BD con grupos de usuarios que comparten los mismos gustos
- Clasificar al usuario dentro de uno de estos grupos
- Recomendar lugares que pueden ser de su agrado

Pipeline

1. Descargar los datos de correos, búsquedas y ubicaciones desde Takeout
2. Procesar cada tipo de dato
3. Vincular los casos de búsquedas y mail a partir del ID-tiempo de los lugares “generadores de diversidad” para buscar coincidencias entre lo que hay en esos lugares y lo que buscó en el explorador o lo que recibió en el mail en un intervalo de tiempo establecido.
4. Utilizar el resultado del minado de texto en los correos y las búsquedas para vincularlo con las ubicaciones de “generadores de diversidad” para identificar qué hay en esos lugares que ayuden a construir un perfil de atributos adicionales.
5. Generar recomendaciones a partir del perfil de lo que cotidianamente consume/visita/hace/busca y con los lugares a los que ha viajado.

Pipeline

- Transformación timestamp a formato fecha/hora/minuto para generar llave útil entre las fuentes de información
- Generar minado de texto en búsquedas



Pipeline

- Transformación timestamp a formato fecha/hora/minuto para generar llave útil entre las fuentes de información
- Extraer de Longitud y Latitud: país, entidad/condado/provincia, ciudad, cp
- Frecuencias para cada nivel de ubicación para ventana de tiempo
- Extraer “zona” más frecuente y definirla como “ciudad residencia”.
- Las ubicaciones fuera de la “ciudad residencia” clasificarlas como “viajes”
- Acotar ubicaciones a los contenidos dentro de “ciudad residencia” y extraer los n puntos más frecuentes para identificarlos como “casa”/”lugar de trabajo”.
- El complemento de las ubicaciones en “ciudad de residencia” clasificarlos como “generadores de diversidad”.



Pipeline

- Transformación de datos de .mbox a formato por definir donde se identifique cada correo y se separe para cada uno cada parte del cuerpo del mismo
- Crear una base de datos en neo4j para representar la red de los correos e identificar las personas con quien se tiene más interacción
- Generar minado de texto en mail, por ejemplo extraer las keywords de los correos y detectar temas en ellos.



Mockup

Perfil

Edad:

Género:

Actividad laboral:

Lugar de residencia:

Lugar de trabajo:

Lugares frecuentados:

Gustos (análisis de búsquedas):

Recomendaciones,

Cerca de su domicilio:

Cerca de su trabajo:

