Bajando datos de Twitter

Usando T-hoarder_kit

Mariluz Congosto @congosto

Índice

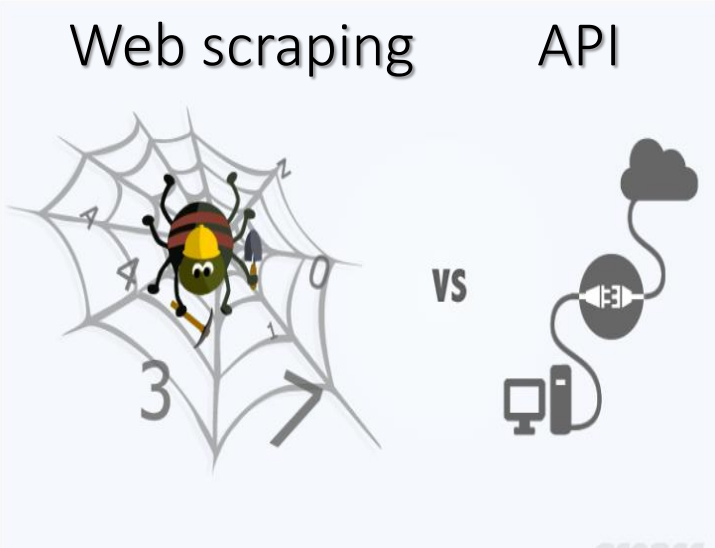
Objetivos del taller

- Comprender el proceso de extracción de datos y su problemática
- Disponer de un kit para aprender y experimentar en la extracción de datos

1. CAPTURA DE DATOS

Captura de datos

- Se puede obtener todo lo publicado en a Web
- No hay rate limit
- Estructura de datos:
  - HTML, XHTML validated
  - JSON, JSONP, XML, Microdata



- Se puede obtener lo que ofrezca la API, a veces más, otras veces menos que en la Web
- Hay rate limit
- Menos volumen de información
- Estructura de datos:
  - JSON

Captura de datos: Autentificación

API

- Autentificación
- Protocolo
- Formatos
- API REST
- API streaming
- Limitaciones
- Consola

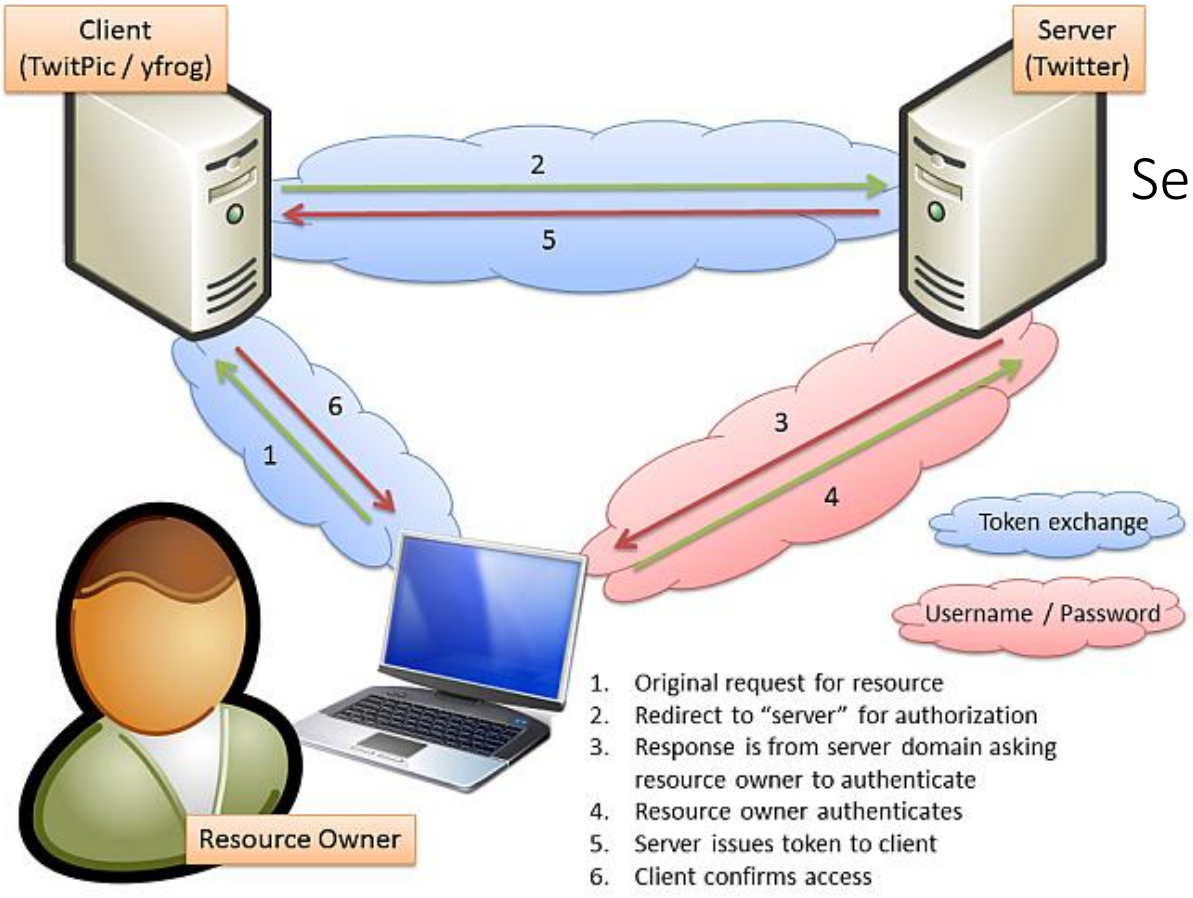
Captura de datos: Autentificación

Captura de datos: Autentificación

Cliente (app)

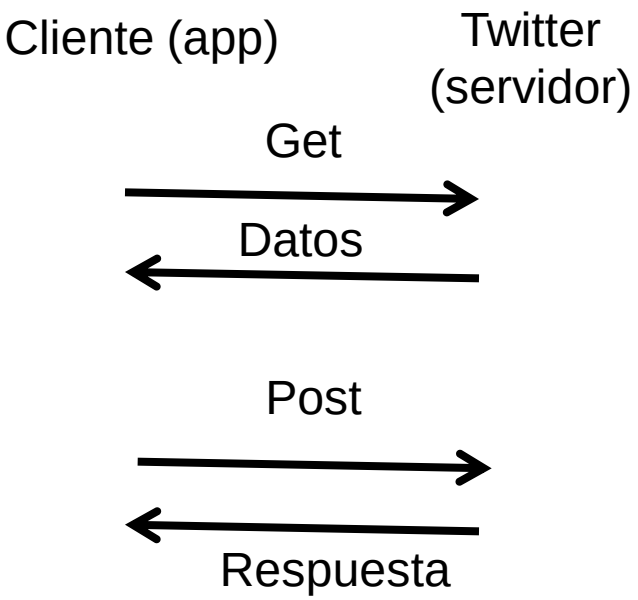
Servidor

Dueño del recurso

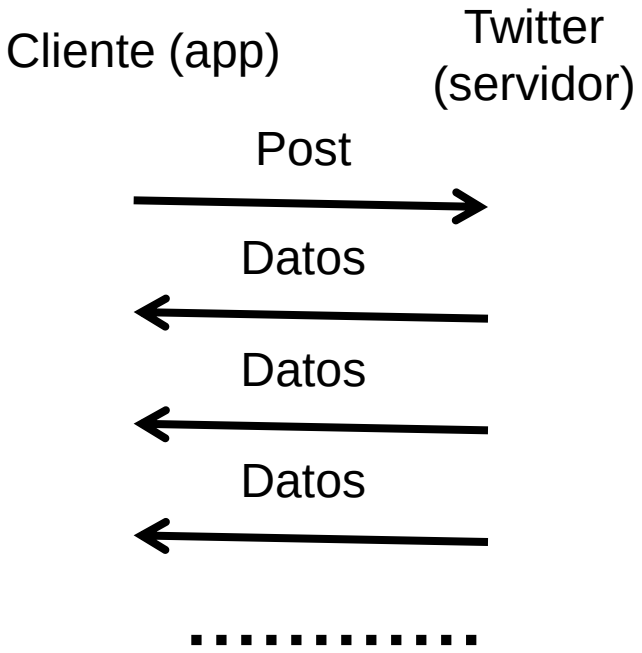


Captura de datos: Protocolos

REST



Streaming



Captura de datos: Formatos

What is JSON?

- JSON stands for **J**ava**S**cript **O**bject **N**otation
- JSON is a lightweight data-interchange format
- JSON is language independent *
- JSON is "self-describing" and easy to understanc

¡¡Menos mal que existen conversores a csv!!

JSON Example

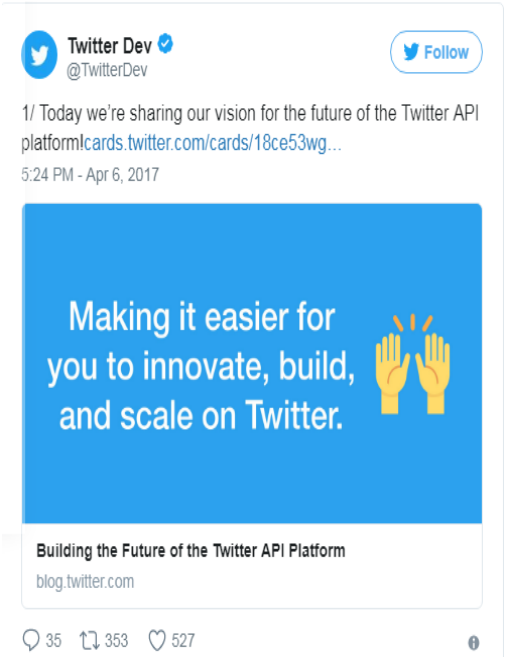
```
{"employees":[
    {"firstName":"John", "lastName":"Doe"},
    {"firstName":"Anna", "lastName":"Smith"},
    {"firstName":"Peter", "lastName":"Jones"}
]}
```

XML Example

```
<employees>
    <employee>
        <firstName>John</firstName> <lastName>Doe</lastName>
    </employee>
    <employee>
        <firstName>Anna</firstName> <lastName>Smith</lastName>
    </employee>
    <employee>
        <firstName>Peter</firstName> <lastName>Jones</lastName>
    </employee>
</employees>
```

http://www.w3schools.com/json/

Captura de datos: Formatos



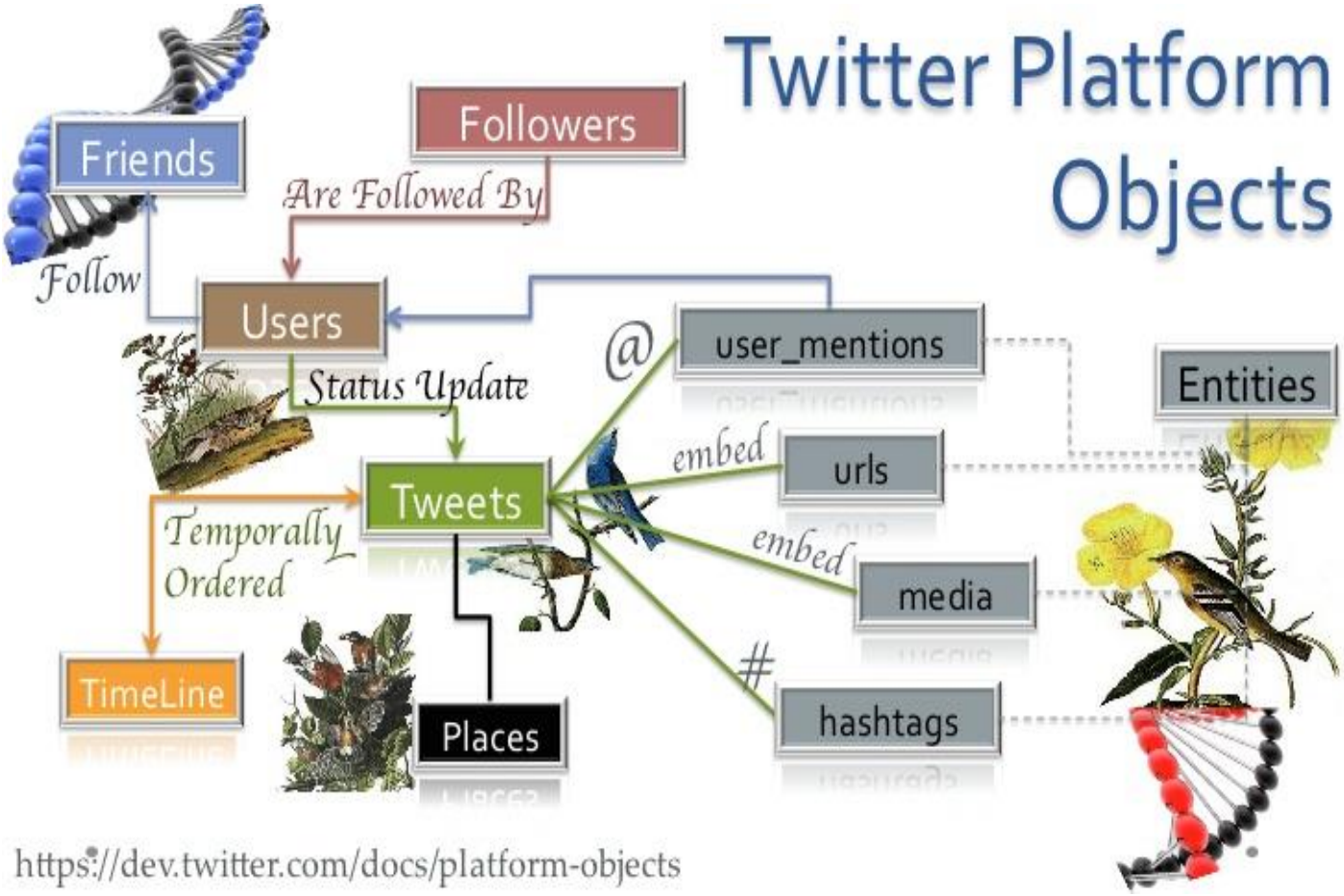

```
{
  "tweet": {
    "created_at": "Thu Apr 06 15:24:15 +0000 2017",
    "id_str": "850006245121695744",
    "text": "1\/ Today we\u2019re sharing our vision for the future of the Twitter API platform!\nht
    "user": {
      "id": 2244994945,
      "name": "Twitter Dev",
      "screen_name": "TwitterDev",
      "location": "Internet",
      "url": "https:\/\/dev.twitter.com\/",
      "description": "Your official source for Twitter Platform news, updates & events. Need technic
    },
    "place": {

    },
    "entities": {
      "hashtags": [

      ],
      "urls": [
```

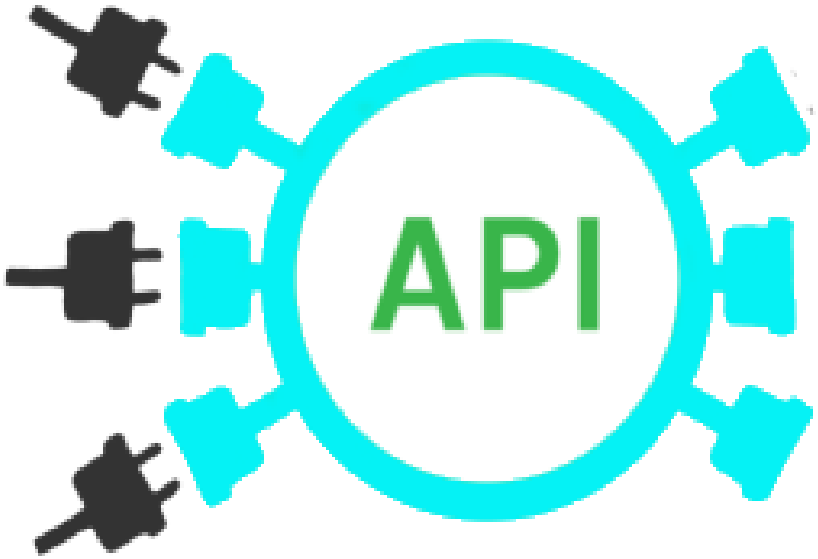

https://developer.twitter.com/en/docs/tweets/data-dictionary/overview/intro-to-tweet-json

Captura de datos: Formatos



Captura de datos: APIs

Twitter APIs



REST API

Permite el acceso al core de los datos de Twitter

Search API

Permite realizar búsquedas

Streaming API

Permite descargar tuits en tiempo real

Rate limit

15-900 — 15 min.

180 — 15 min.

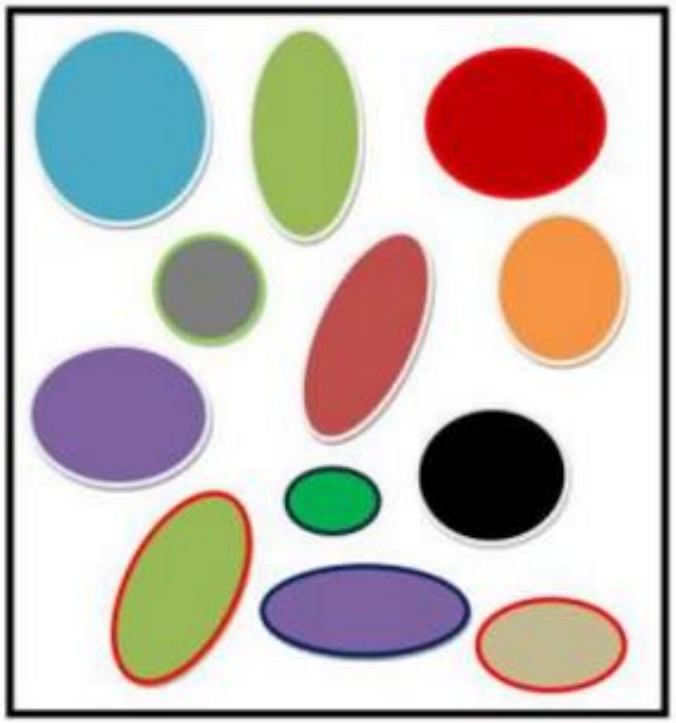
50 — 1 seg.

Captura de datos: API REST

Familias

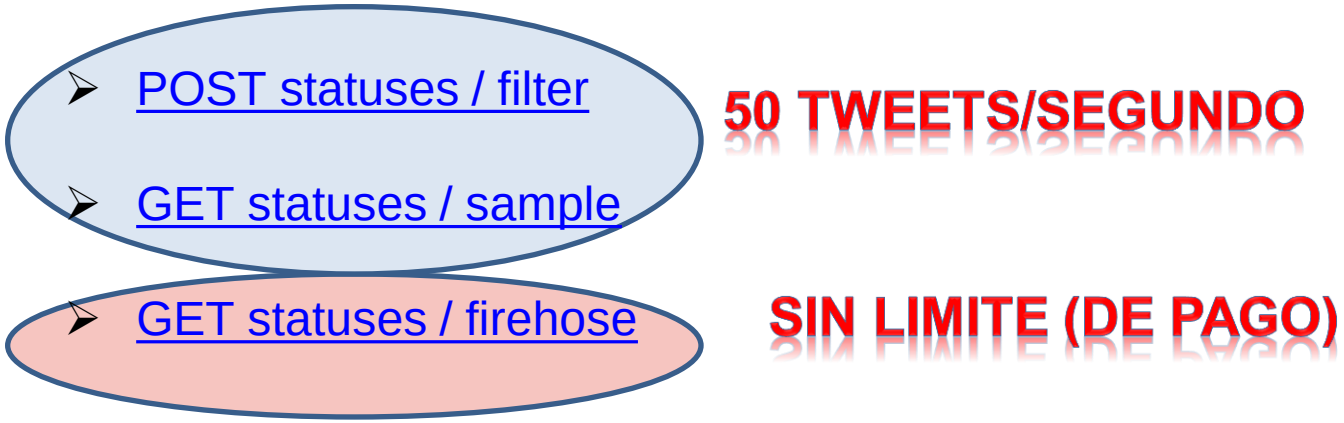
- **Search API**
- **Favoritos**
- **Statuses**
- **Users**
- **Followers**
- **Friends**
- **Friendships**
- **Lists**
- **Trends**
- **Help**

Métodos



Captura de datos: API Streaming

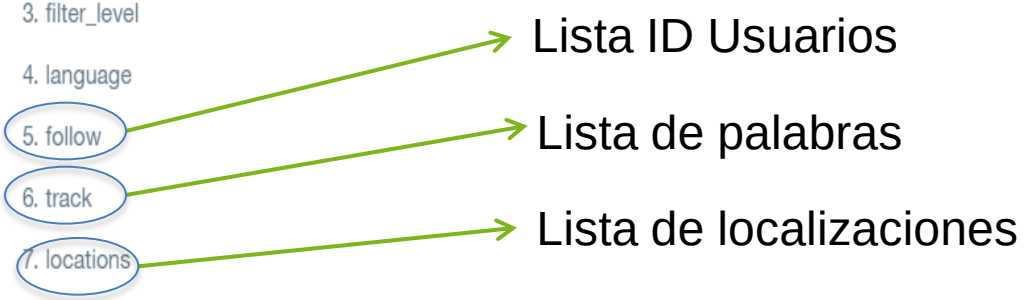
Rate limit



Captura de datos: API Streaming

Basic Streaming API request parameters

1. delimited
2. stall_warnings
3. filter_level
4. language
5. follow
6. track
7. locations
8. count
9. with
10. replies
11. stringify_friend_id



https://developer.twitter.com/en/docs/tweets/filter-realtime/guides/basic-stream-parameters

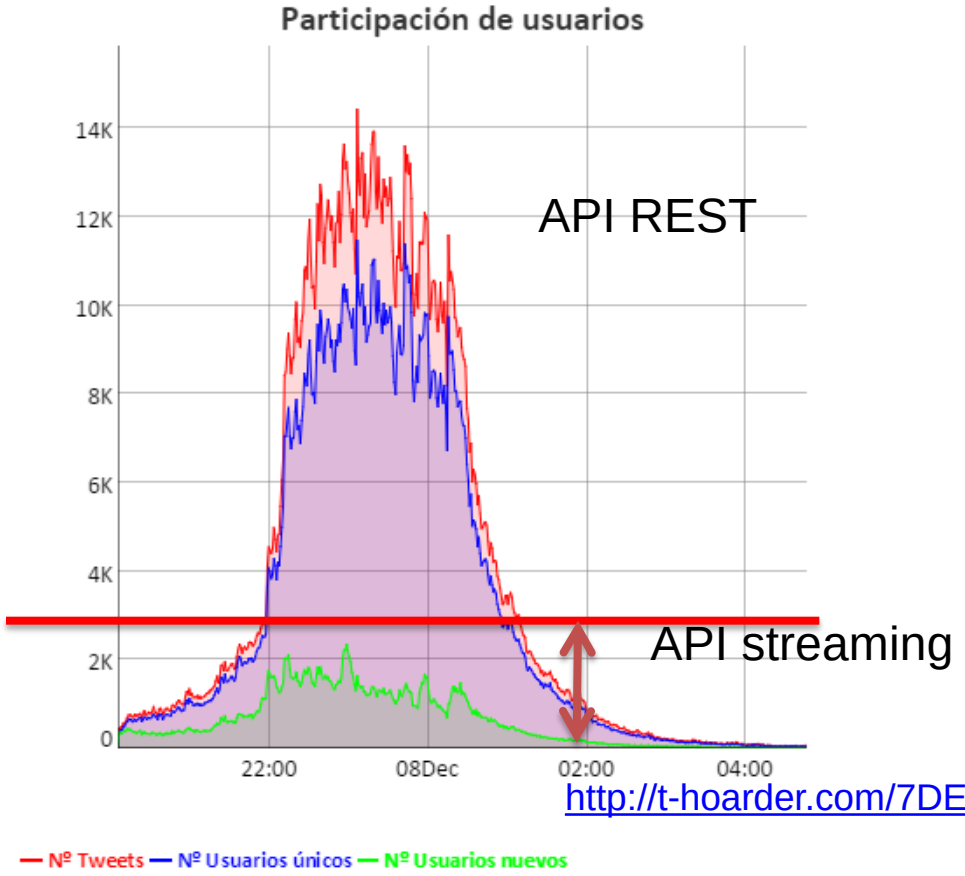
Captura de datos: limitaciones

1. Cantidad de información
   - REST: número máximo de consultas cada 15 minutos
   - Streaming: máximo de 50 tuits por segundo
2. Historial
   - REST: una semana de antigüedad
   - Streaming: ninguna antigüedad

Captura de datos: limitaciones

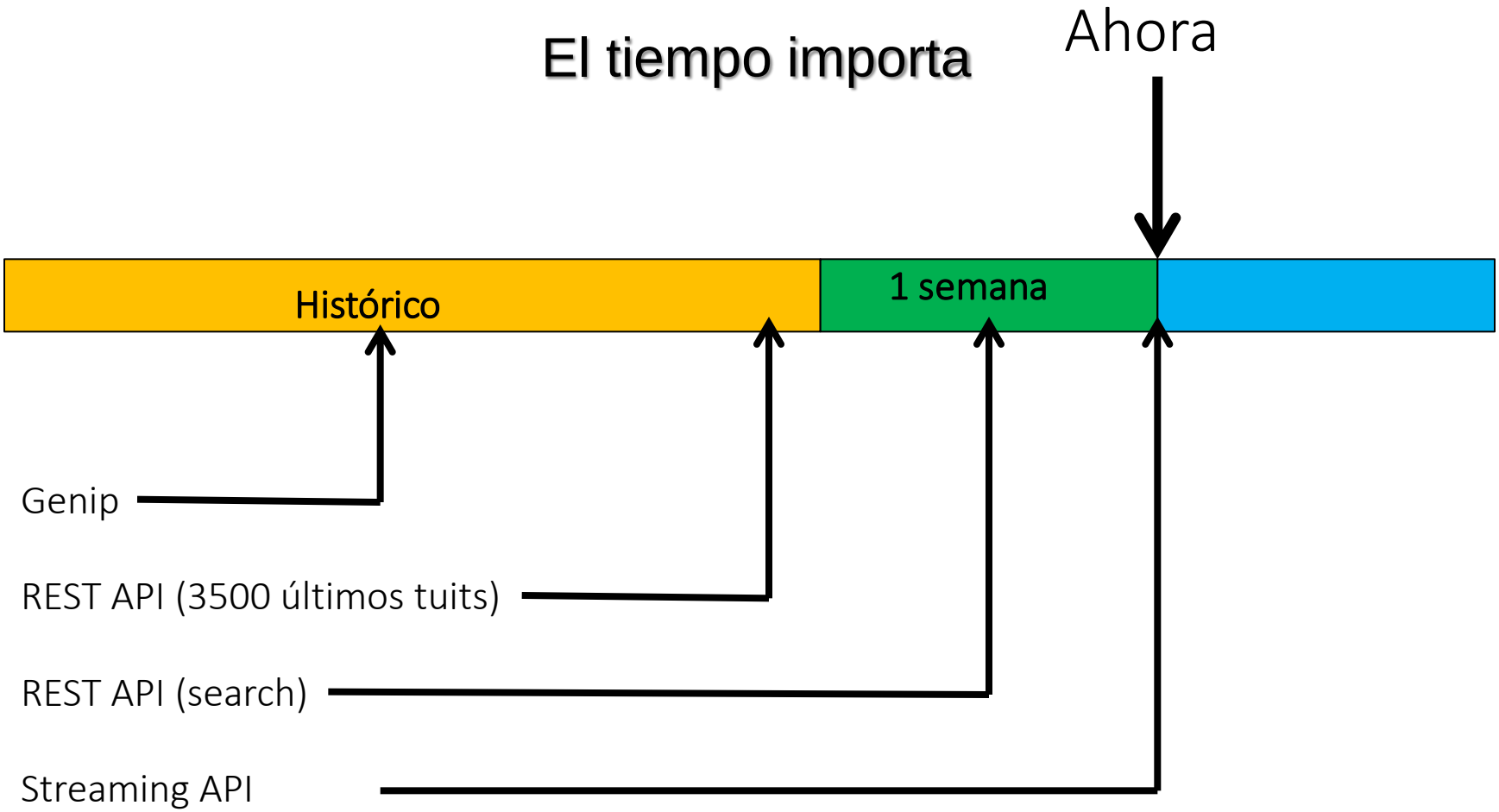
La frecuencia importa: el efecto meseta

#7DEldebateDecisivo



http://t-hoarder.com/7DElDebateDecisivo/

Captura de datos: limitaciones



2. TALLER DE USO DE LA API DE TWITTER

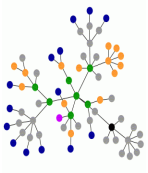
Taller de uso de la API de Twitter

Entorno de trabajo

Autentificar

Obtener información de usuarios

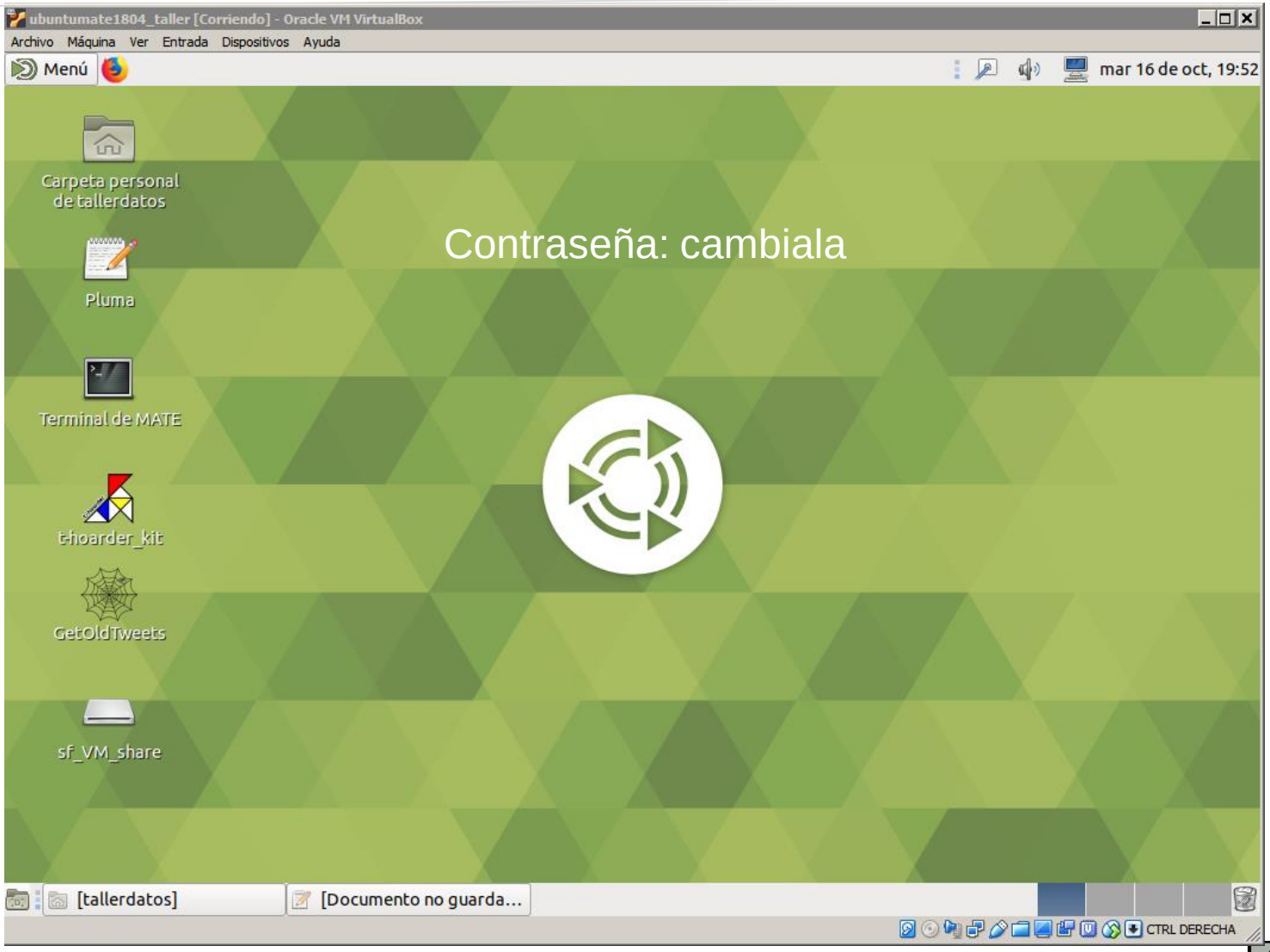
Obtener tuits



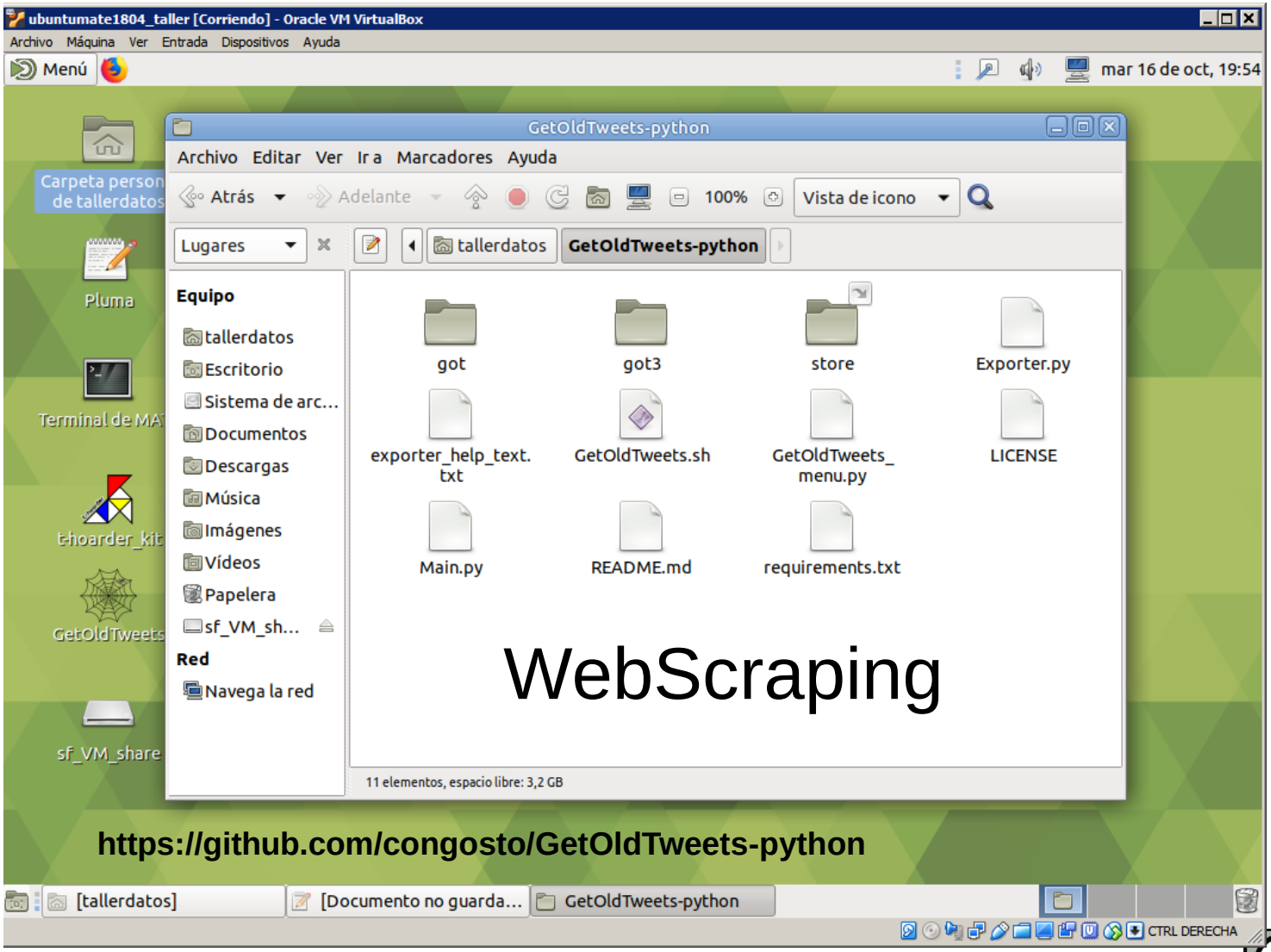
Obtener relaciones

ENTORNO DE TRABAJO

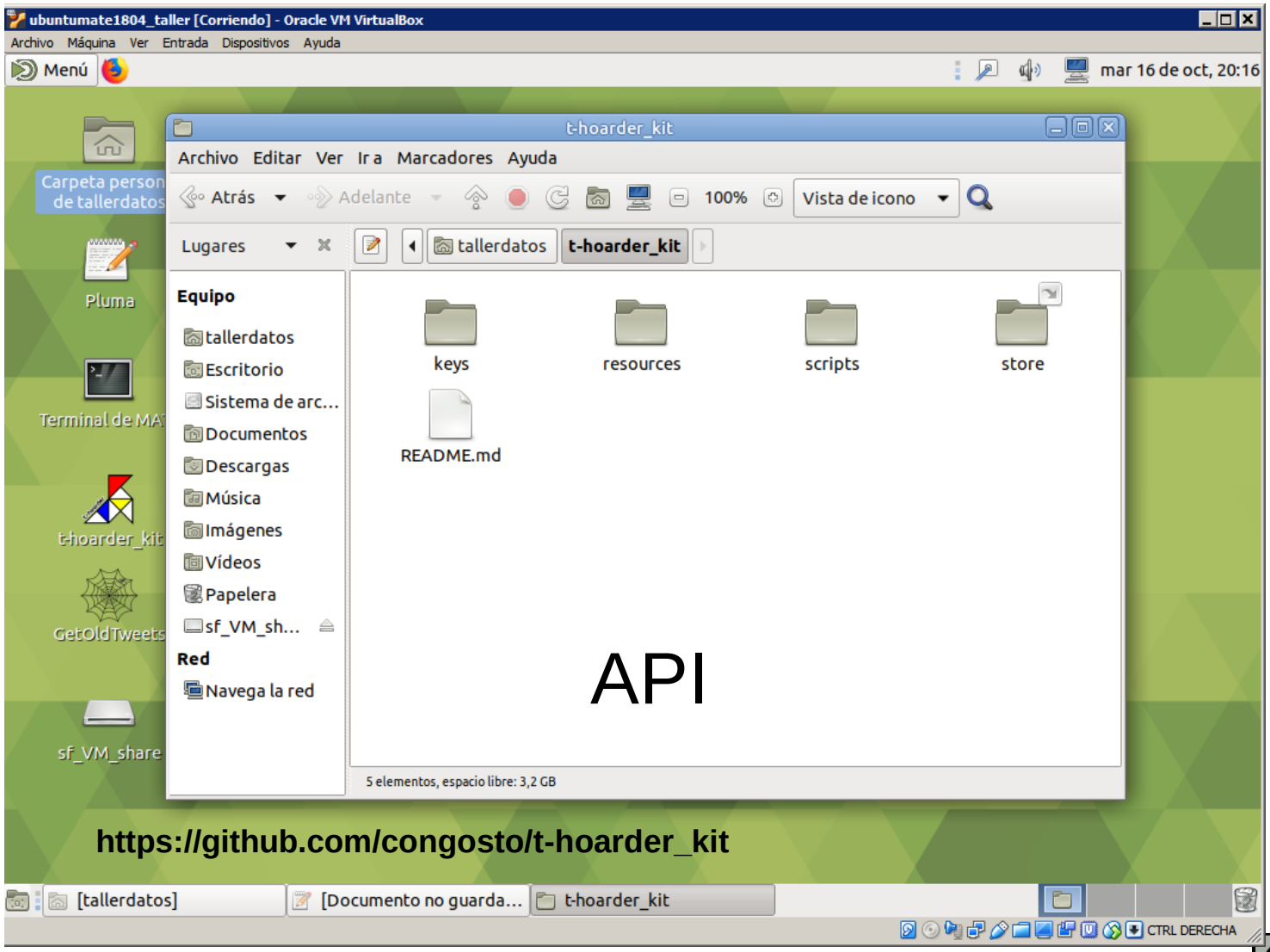
Entorno de trabajo



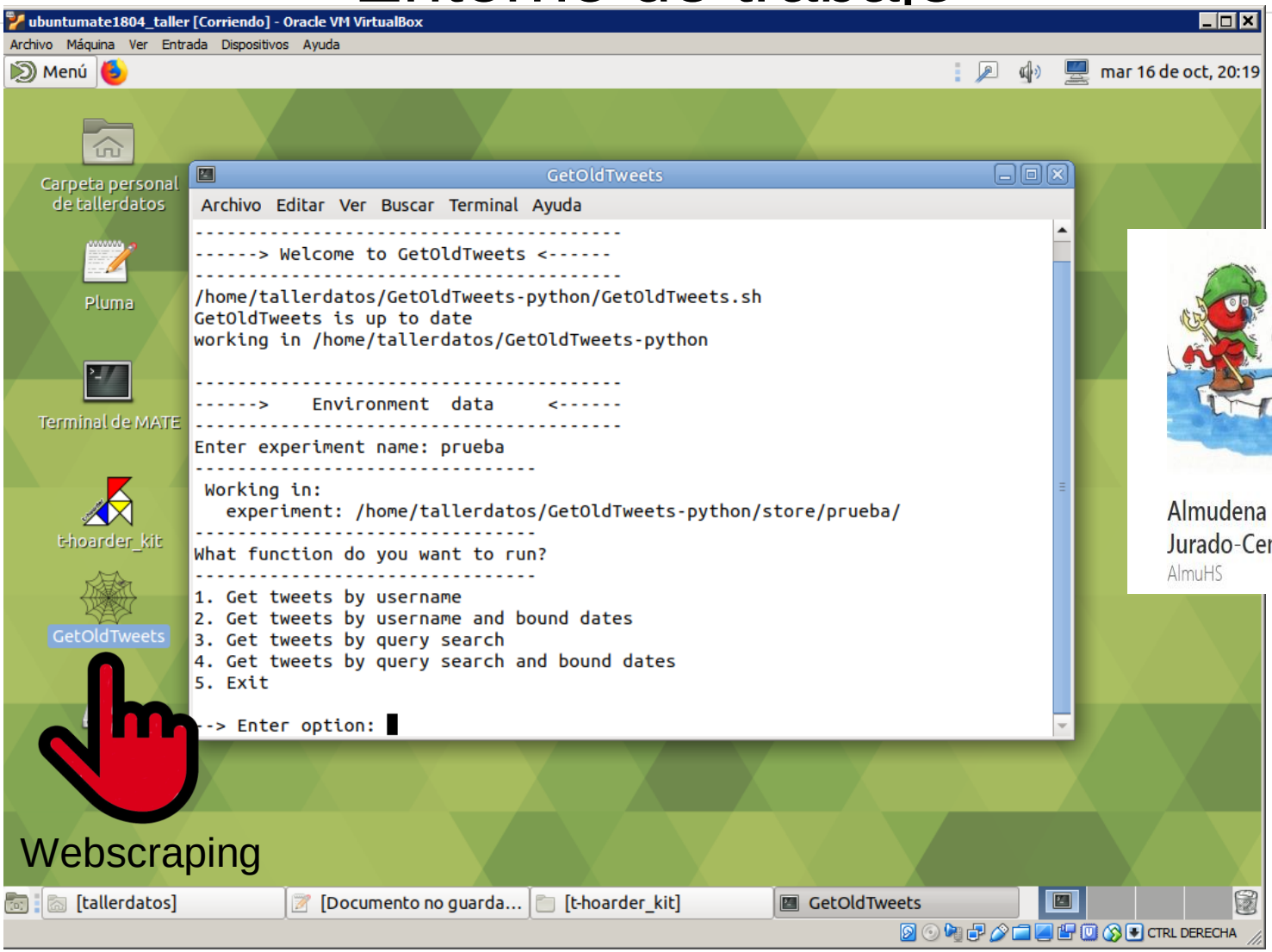
Entorno de trabajo



Entorno de trabajo

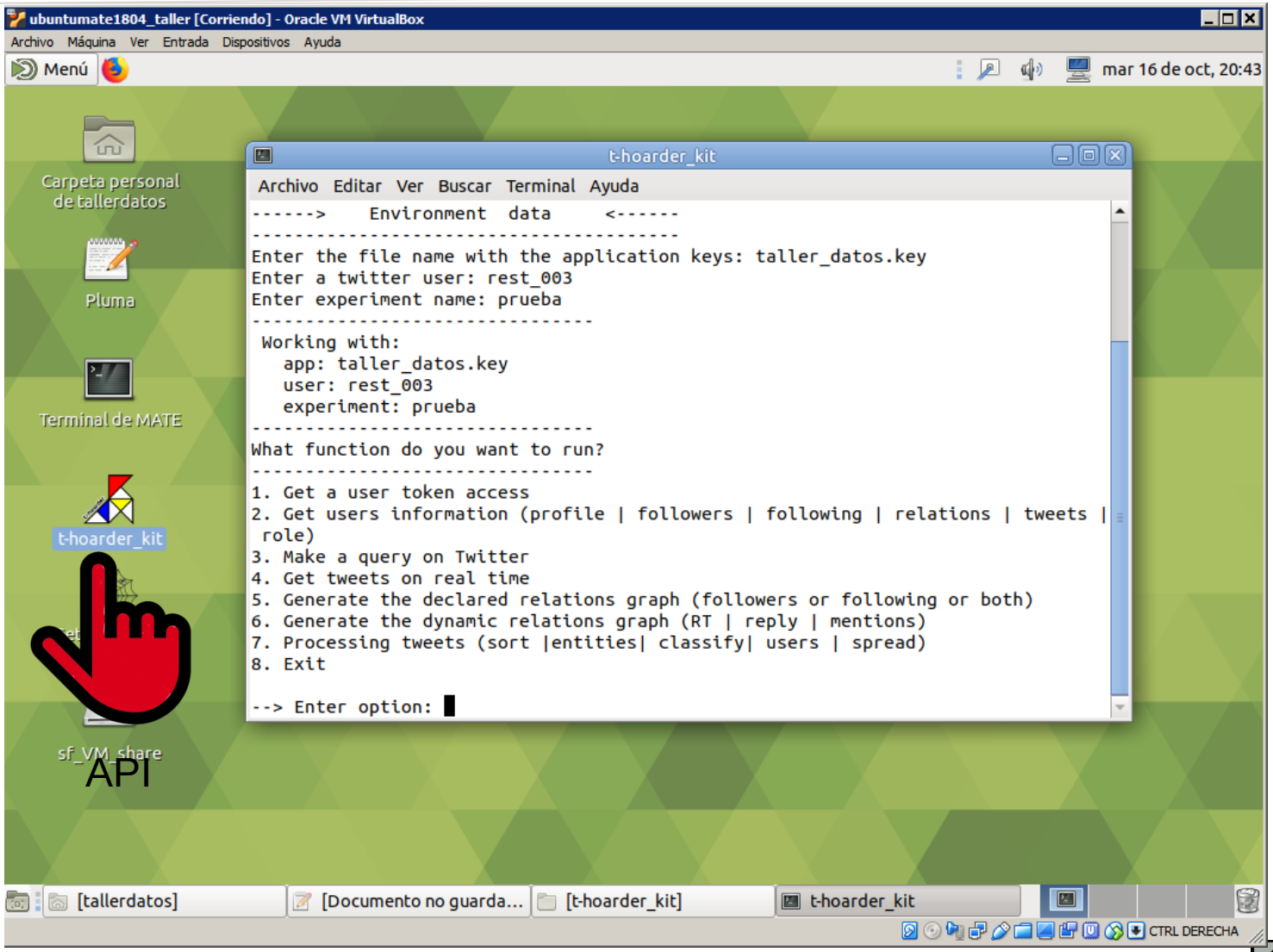


Entorno de trabajo



Almudena Garcia Jurado-Centurion
AlmuHS

Entorno de trabajo

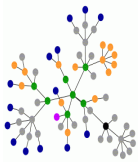


Descarga de datos

Autentificar

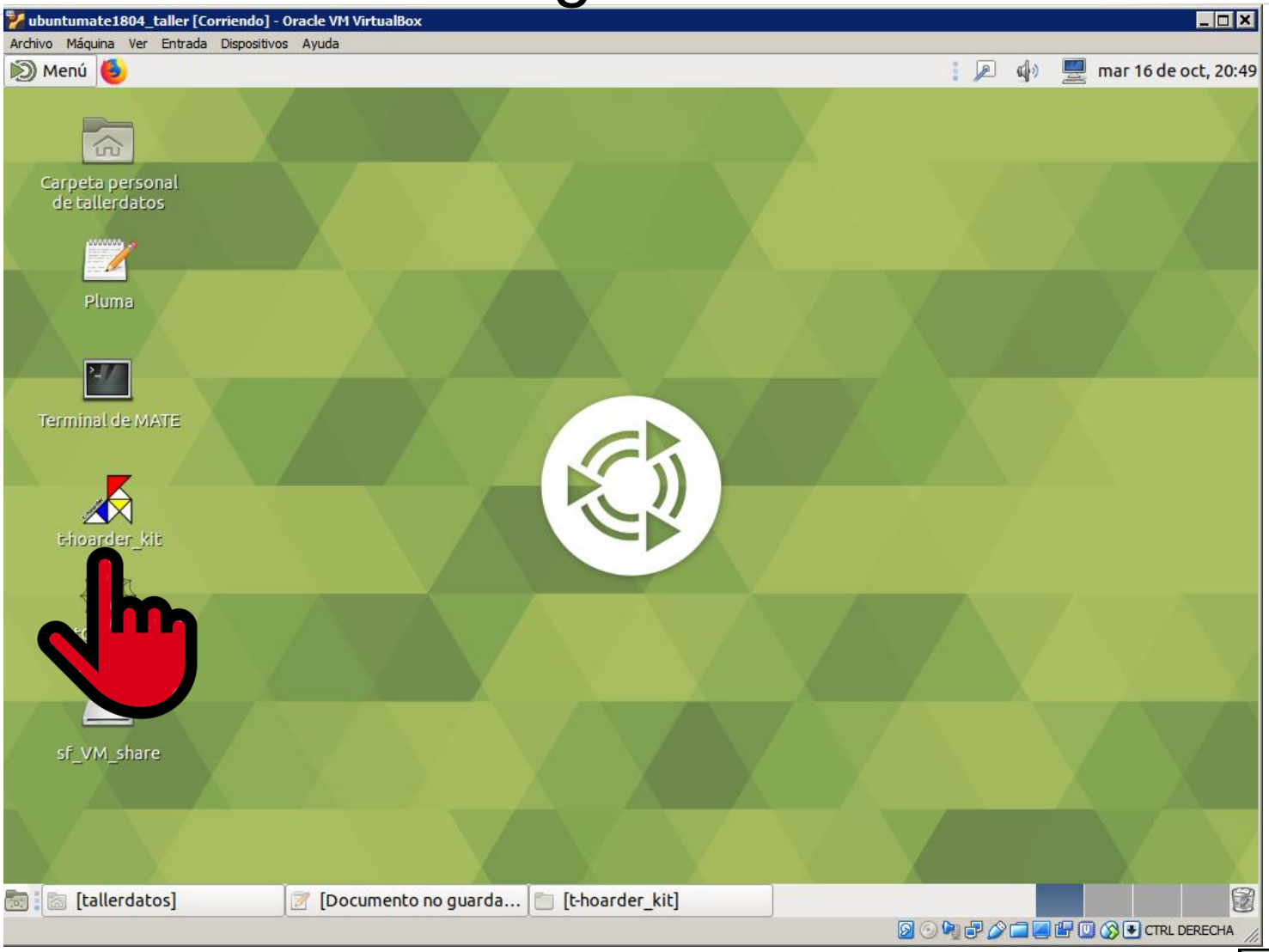
Obtener información de usuarios

Obtener tuits



Obtener relaciones

Descarga de datos



Datos de contexto

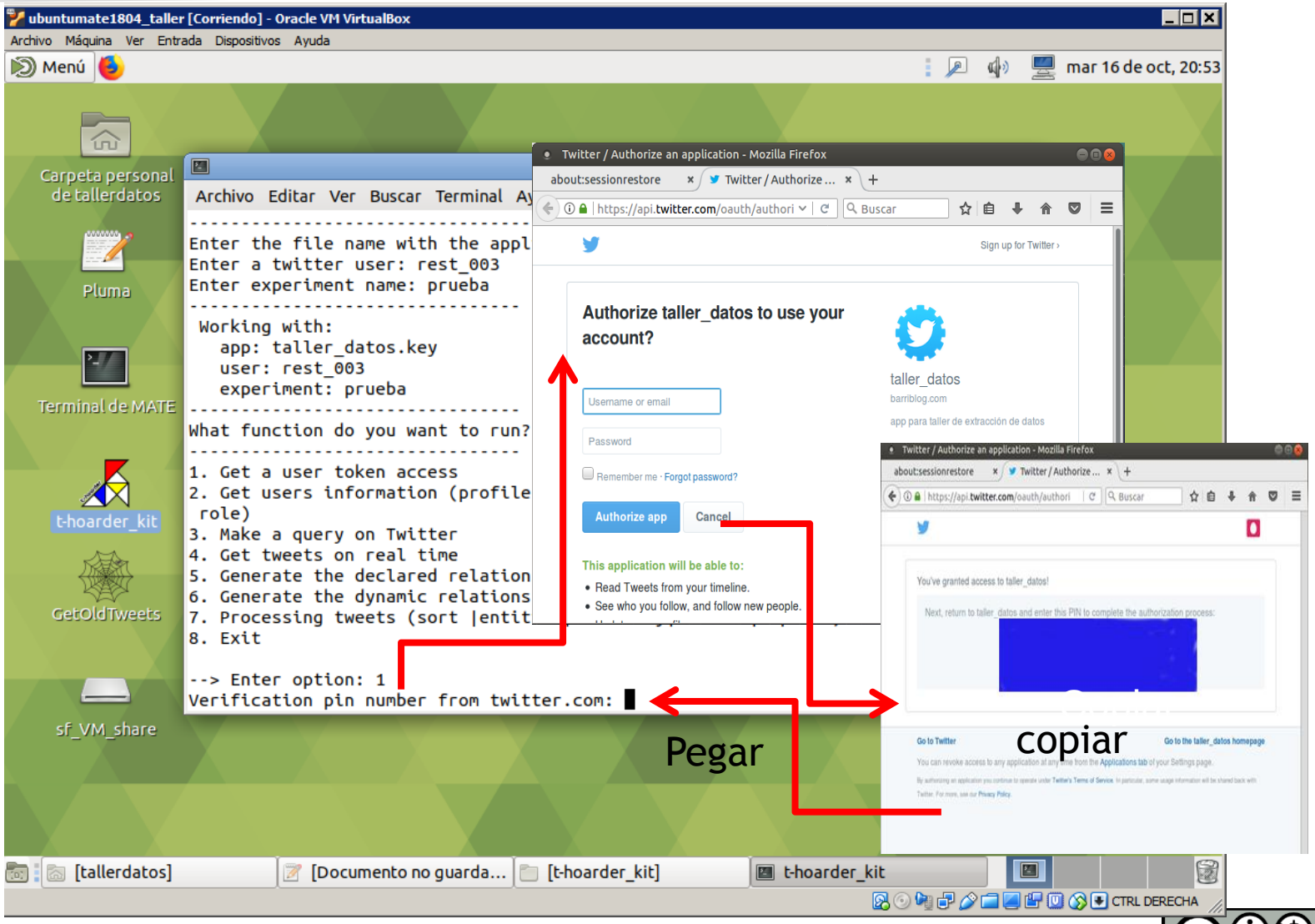
App para acceder -> taller_datos.key

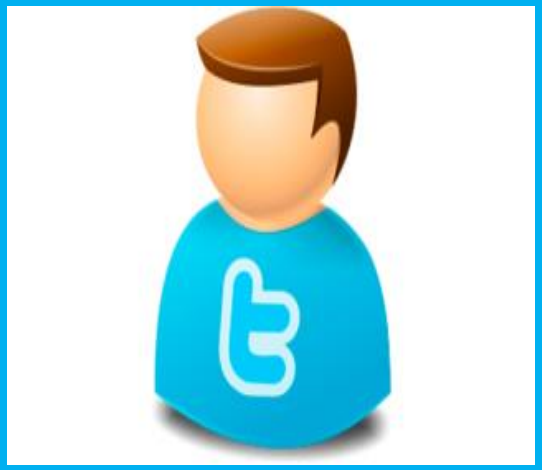
Usuario: cada uno el suyo

Experimento: directorio donde se dejarán los datos

Se introduce la principio y todos las peticiones utilizarán ese contexto

Descarga de datos





OBTENER INFORMACIÓN DE USUARIOS

Información de usuarios

Perfil

Perfiles de sus seguidores

Perfiles de sus seguidos

Perfiles de sus contactos

Tweets publicados

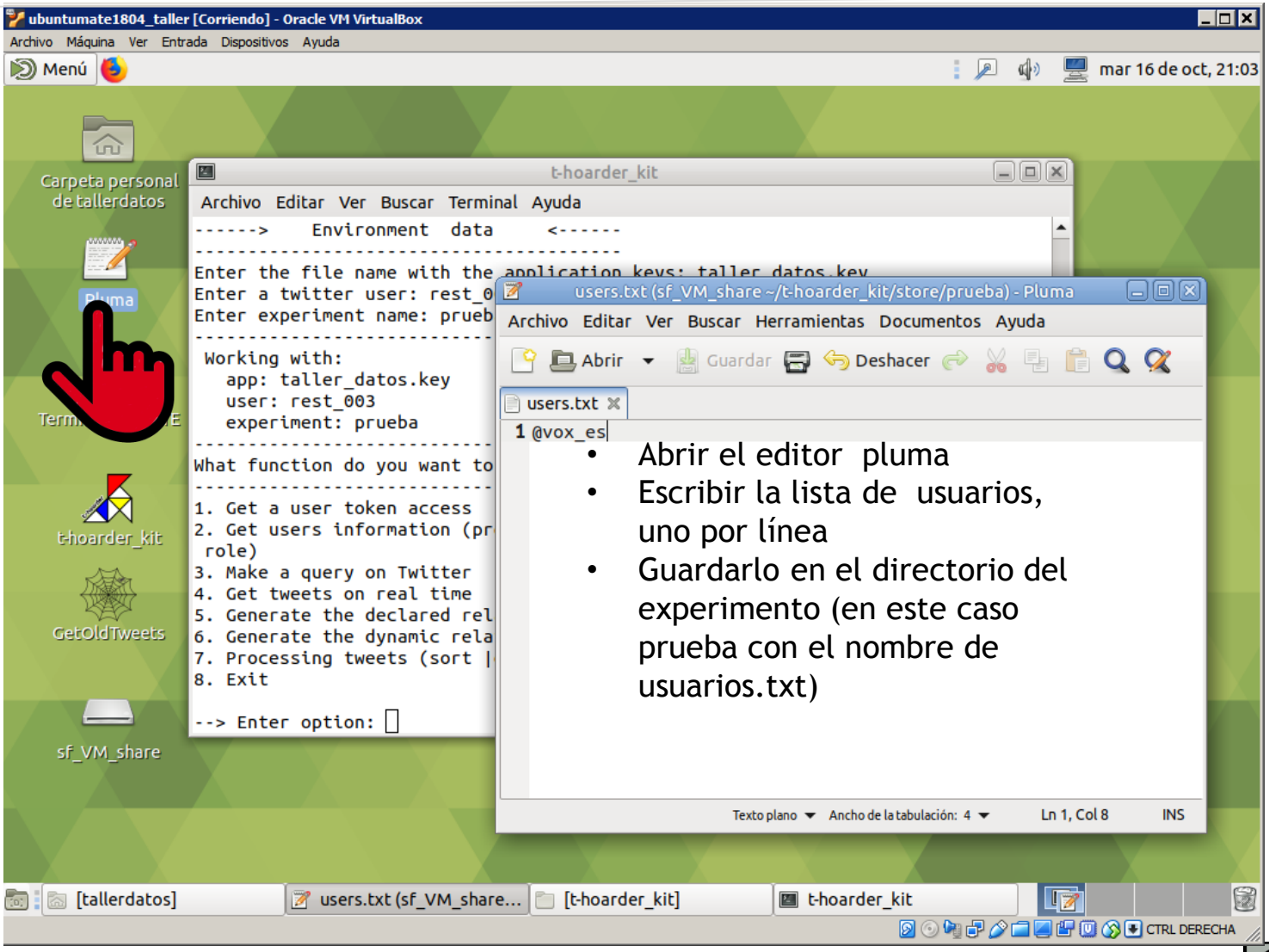
Role

Conexiones

Información de usuarios

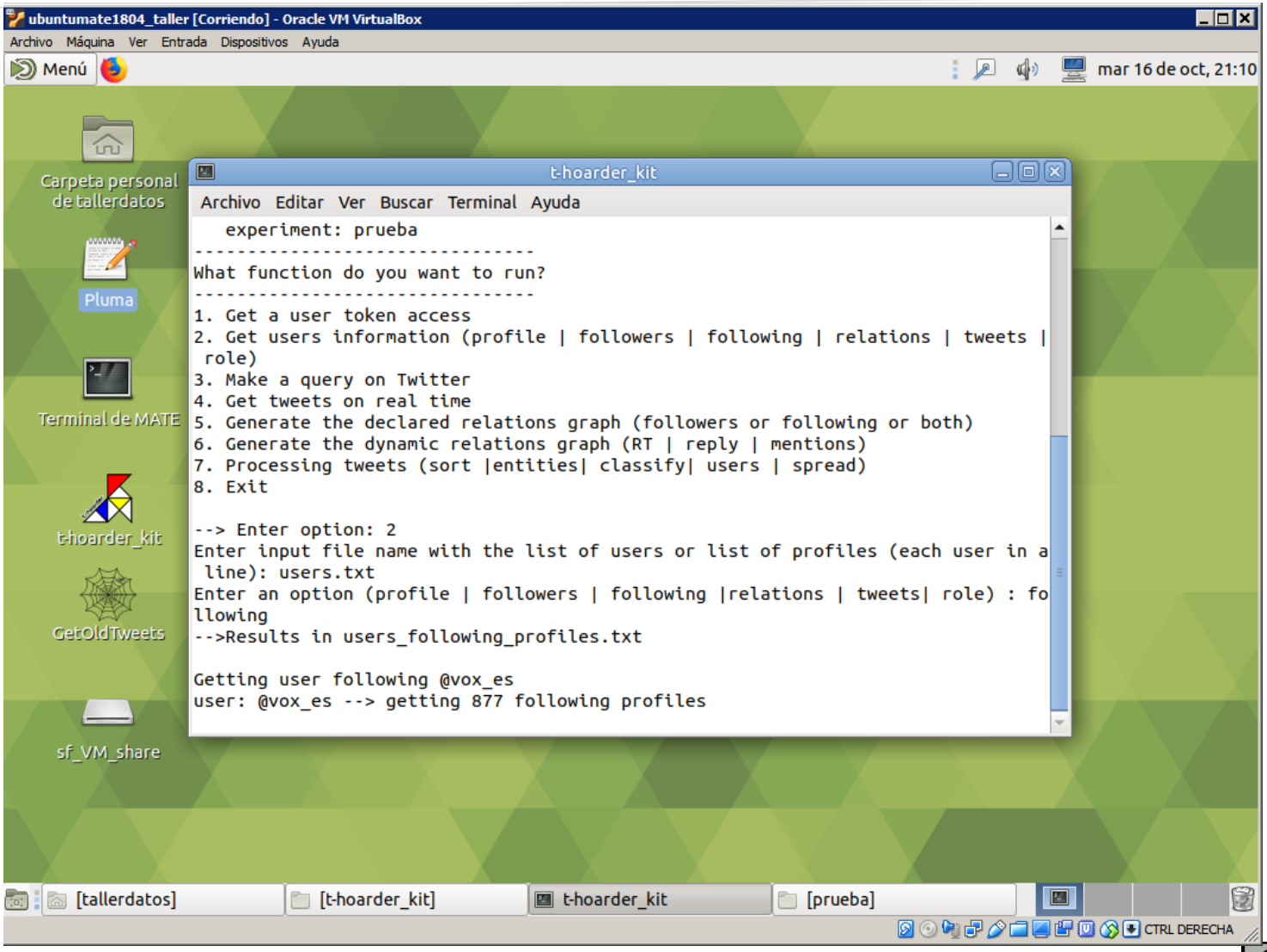
Operación	Método	Limitaciones
--profiles	GET users/show	3.600 perfiles /hora
--followers	GET followers/list	12.000 perfiles/hora
--following	GET friends/list	12.000 perfiles/hora
--relations	GET followers/list GET friends/list	12.000 perfiles/hora
--tweets	GET statuses/user_timeline	720.000 tweets/hora

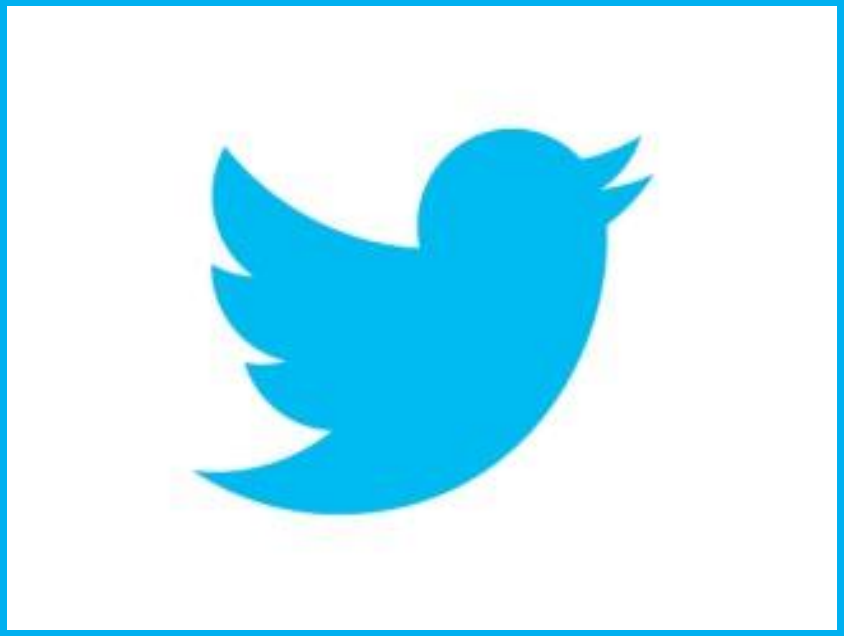
Información de usuarios



- Abrir el editor pluma
- Escribir la lista de usuarios, uno por línea
- Guardarlo en el directorio del experimento (en este caso prueba con el nombre de usuarios.txt)

Información de usuarios





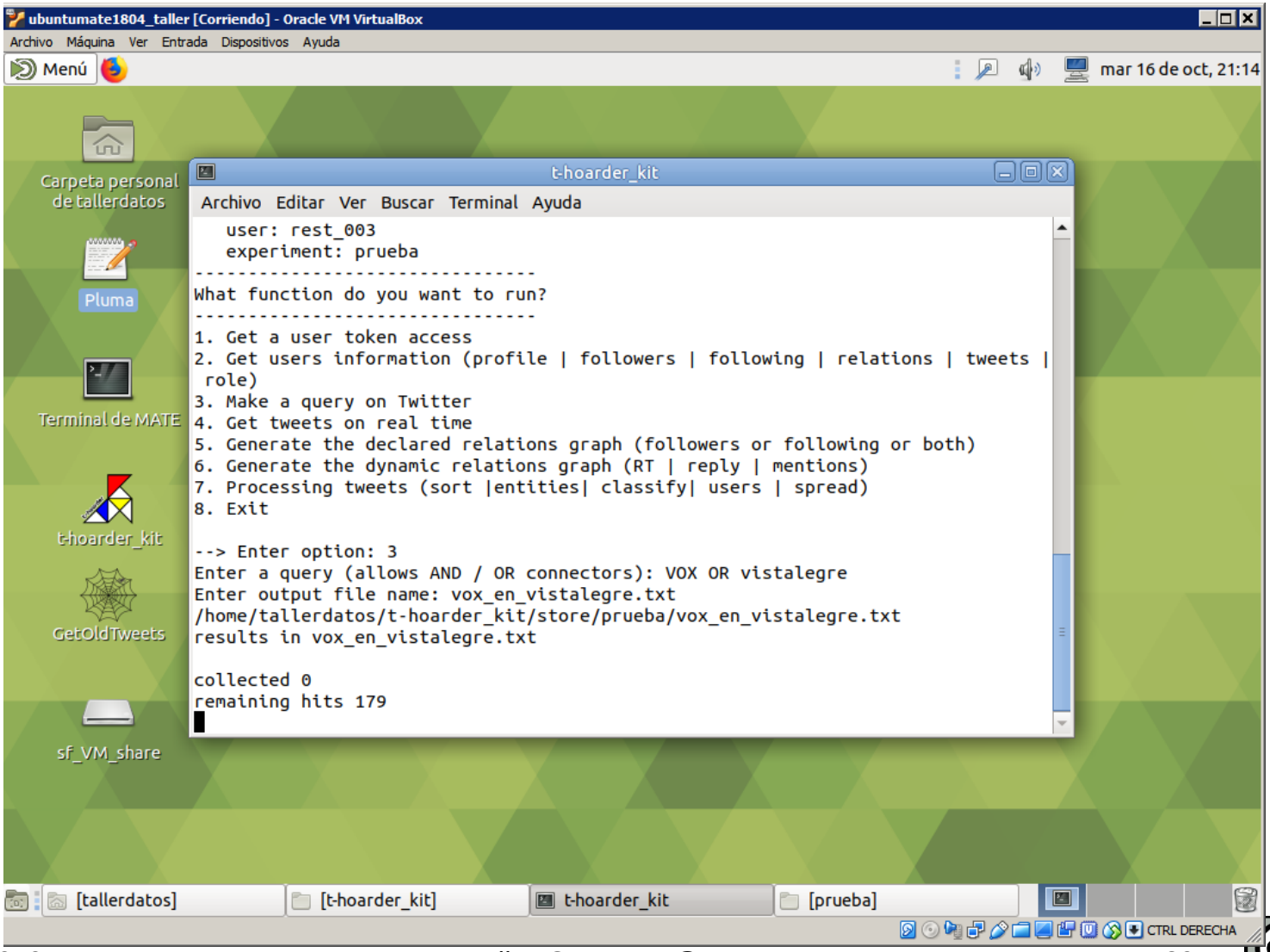
OBTENER TUITS

Obtener tuits

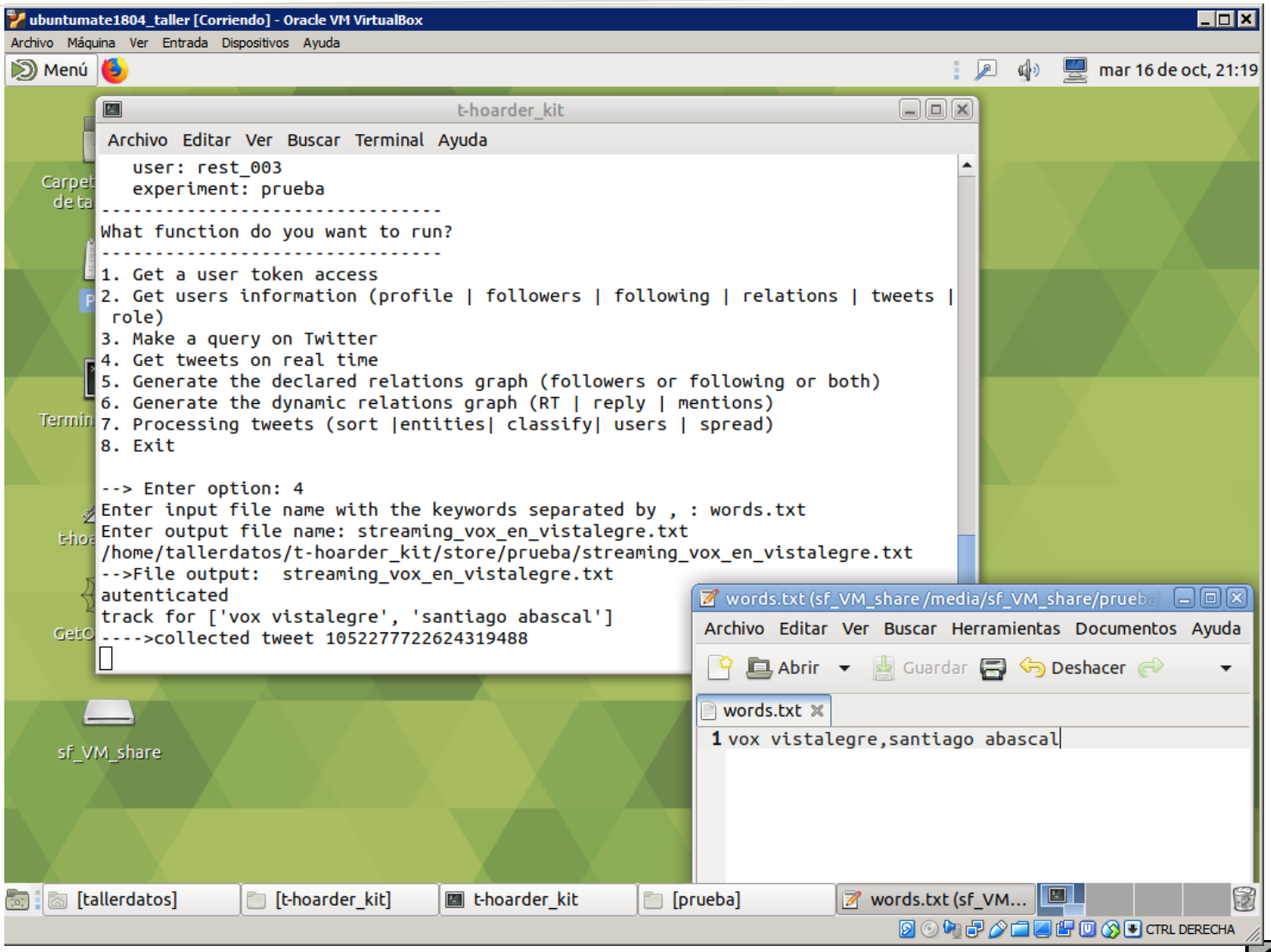
Operación	Método	Limitación
Buscar tweets	GET search/tweets	72.000 tweets /hora
Bajar tweets en tiempo real	POST statuses_filter	Máximo de 180.000 tweets/hora

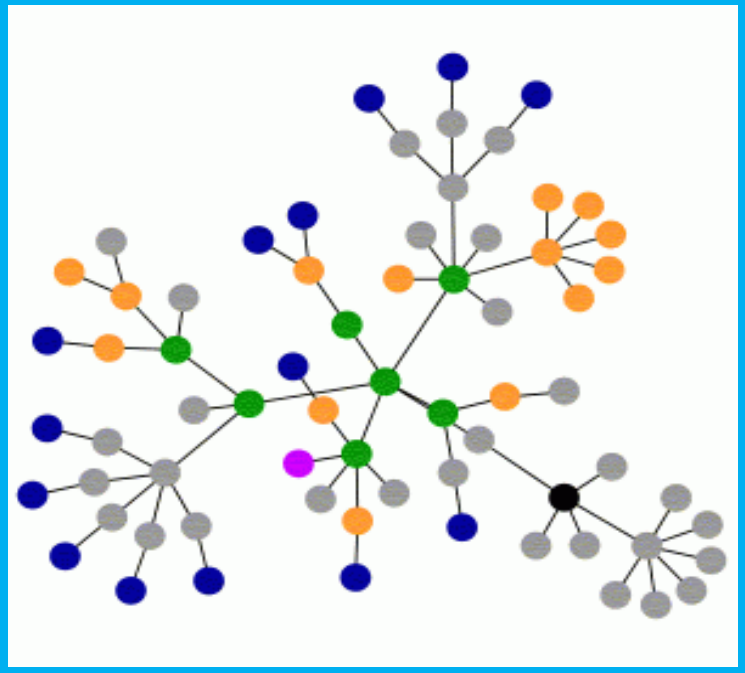
https://twitter.com/search-advanced?lang=es

Obtener tuits



Obtener tuits



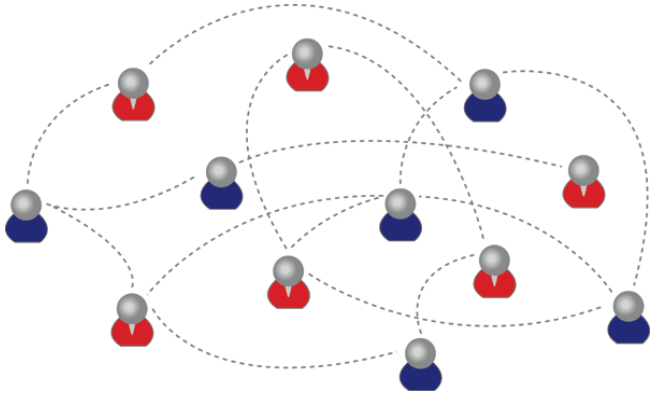


OBTENER RELACIONES

Obtener relaciones

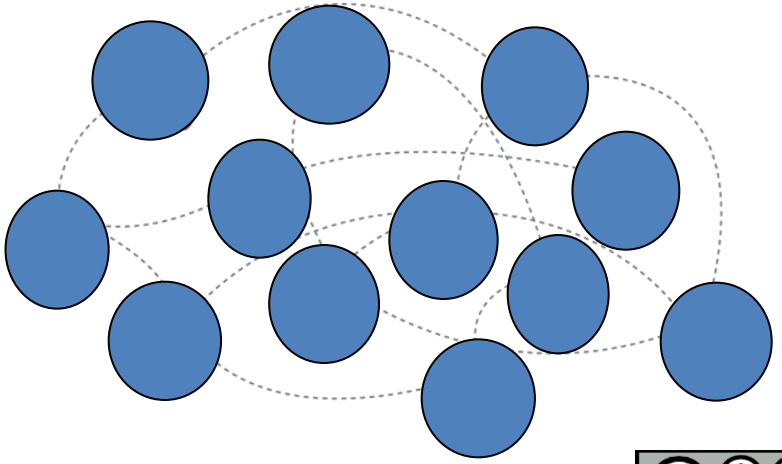
Con teoría de **grafos**, que modela:

- Individuos como **nodos**
- Relaciones como **aristas**



Un **grafo** es una abstracción que representa una red, donde:

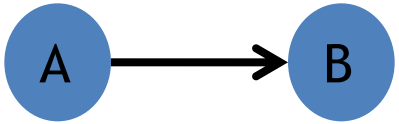
- Un conjunto de **nodos** o **vértices** está conectado mediante **aristas** o **enlaces**



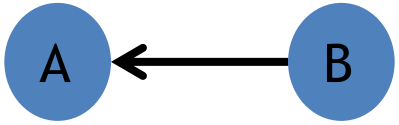
Obtener relaciones

Relaciones declaradas

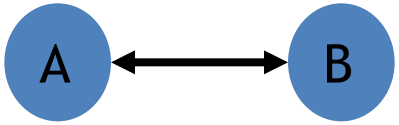
A sigue a B



A es seguido por B

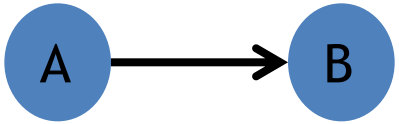


A y B se siguen mutuamente



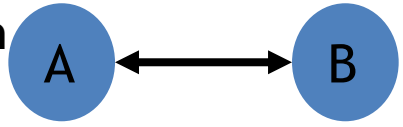
Relaciones dinámicas

A retuitea a B



A es retuiteado por B

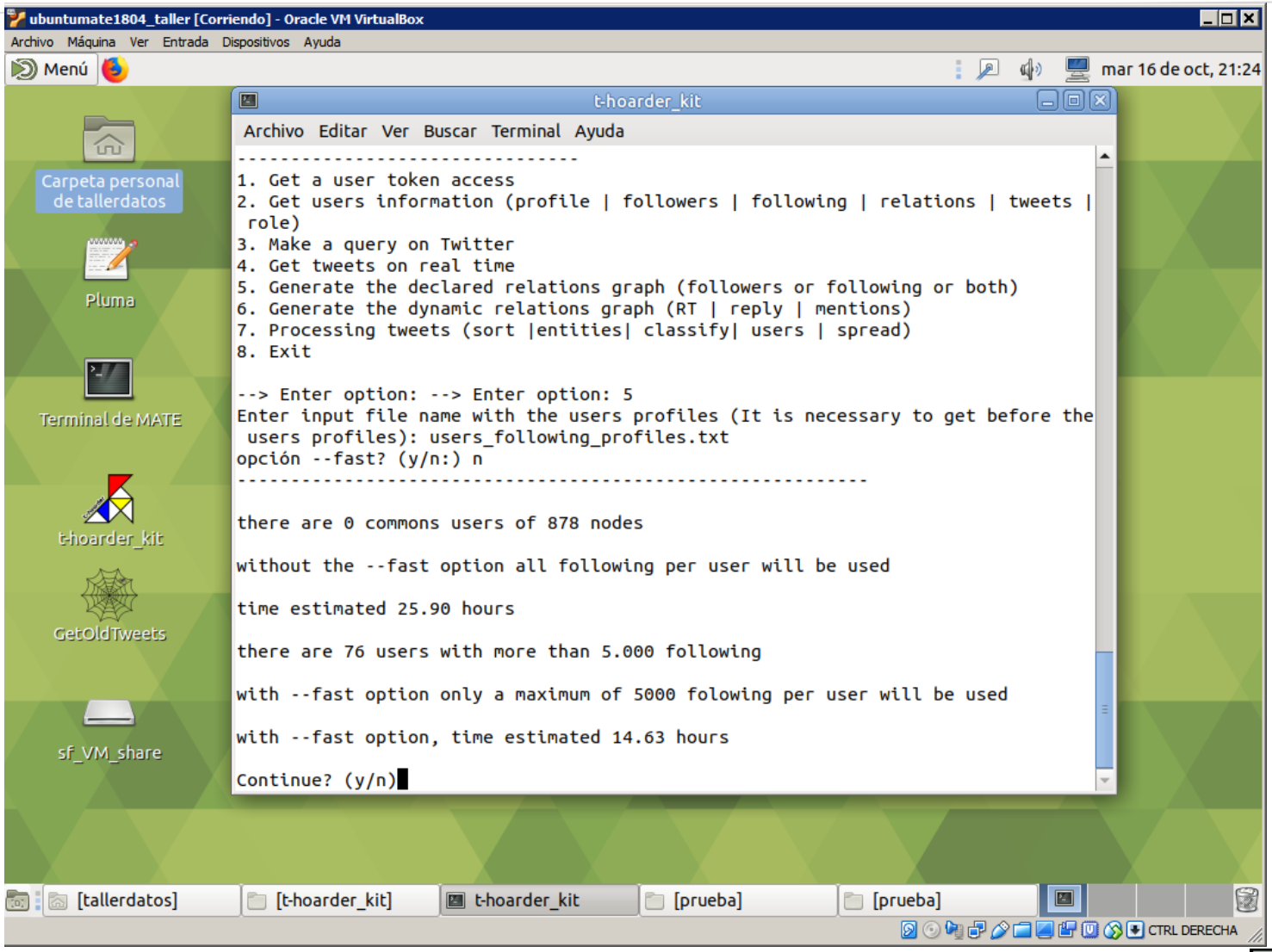
A y B se retuitean mutuamente



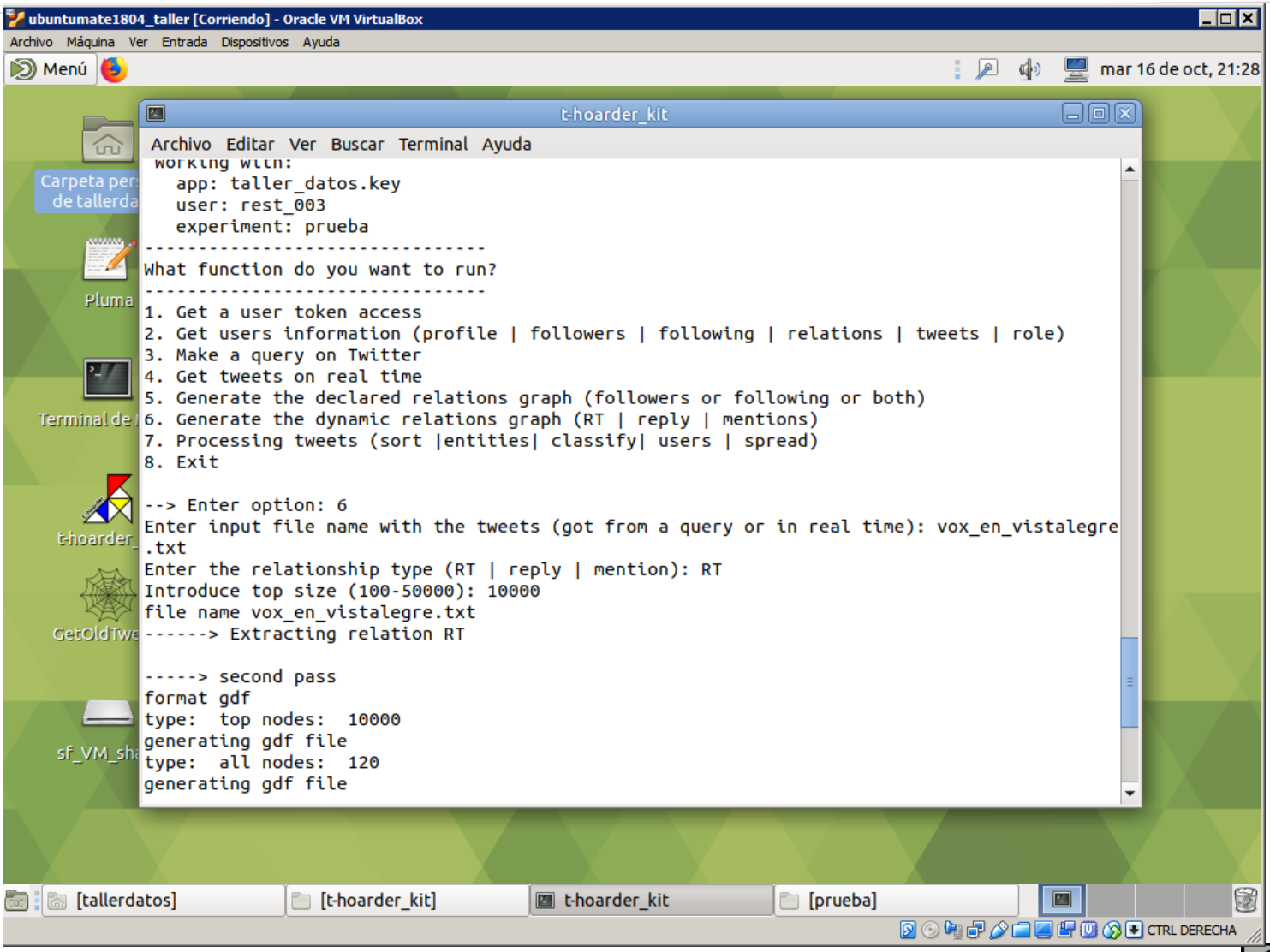
Obtener relaciones

Operación	Método	Limitación
Relaciones declaradas	GET followers/ids GET friends/ids	60 peticiones hora Máximo 300.000 ids /hora (5.000 Ids por petición) Con –fast 60 conexiones de usuarios /hora
Relaciones dinámicas	No necesita la API-	-

Obtener relaciones



Obtener relaciones



Resumen

- Comprender el proceso de extracción de datos y su problemática
- Asumir que obtener los datos (gratis), tiene limitaciones de volumen y temporales
- Disponer de un kit para experimentar y profundizar en estos conceptos
- Posibilidad de mejorar o ampliar los scripts de t-hoarder_kit