Modelo de predicción usando Reeb

Mitsiu Alejandro Carreño Sarabia - E23S-18014

Introducción

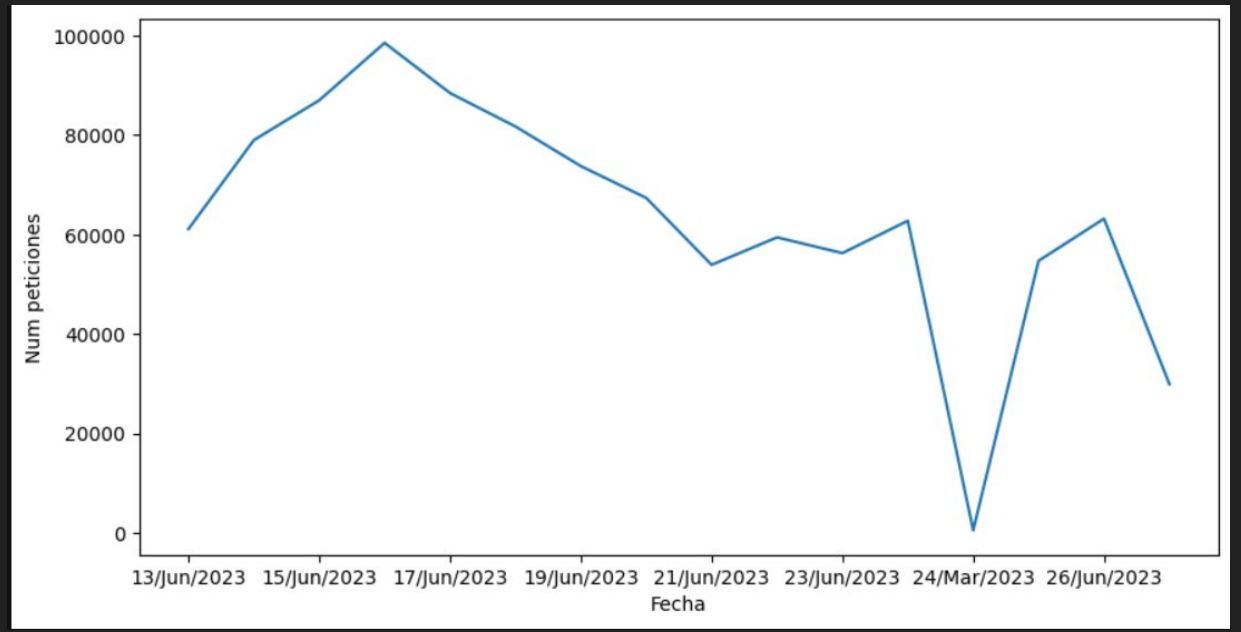
Se tiene acceso a un servidor alojando **~40 dominios web**

[servicios.ieec.mx, sii.upa.edu.mx, www.aguasvoyalcentro.com, stage.aguasvoyalcentro.com, megawatt.com.mx, …]

Cada dominio está enfocado en temáticas distintas, [comercial, educación, administración de recursos…] por lo que **la naturaleza de su demanda varía**.

Introducción

Del cual se recolectó **1M de peticiones**, en un periodo de **16 días**, (~65k peticiones al día)



Introducción

Por cada petición se obtienen **21 características** entre las que destacan

	remote_addr	date_time	req_uri	status	body_bytes_sent
0	185.213.174.190	[27/Jun/2023:07:12:12 -0600]	/	502.0	575.0
1	185.213.174.190	[27/Jun/2023:07:12:12 -0600]	/index.php?s=/index/think%5Capp/invokeMethod&method[0]=think%5Cview%5Cdriver%5CPhp&method[1]=display&vars[0]=%3C?php%20echo%20md5(%271f3870be274f6c49b3e31a0c6728957f%27);	502.0	575.0
2	185.213.174.190	[27/Jun/2023:07:12:13 -0600]	/index.php?s=/admin/think%5Capp/invokeMethod&method[0]=think%5Cview%5Cdriver%5CPhp&method[1]=display&vars[0]=%3C?php%20echo%20md5(%271f3870be274f6c49b3e31a0c6728957f%27);	502.0	575.0
3	185.213.174.190	[27/Jun/2023:07:12:14 -0600]	/index.php?s=/api/think%5Capp/invokeMethod&method[0]=think%5Cview%5Cdriver%5CPhp&method[1]=display&vars[0]=%3C?php%20echo%20md5(%271f3870be274f6c49b3e31a0c6728957f%27);	502.0	575.0
4	185.213.174.190	[27/Jun/2023:07:12:14 -0600]	/index.php?s=/home/think%5Capp/invokeMethod&method[0]=think%5Cview%5Cdriver%5CPhp&method[1]=display&vars[0]=%3C?php%20echo%20md5(%271f3870be274f6c49b3e31a0c6728957f%27);	502.0	575.0

Problemática

Mucho se ha desarrollado en términos de escalabilidad de infraestructura así como adopción de soluciones distribuidas para dar servicio a la creciente demanda (por ello podemos manejar 65k peticiones al día).

La cantidad de **información generada es tan grande** que un análisis manual no es viable.

Problemática

Pero el análisis de peticiones ofrece grandes beneficios:

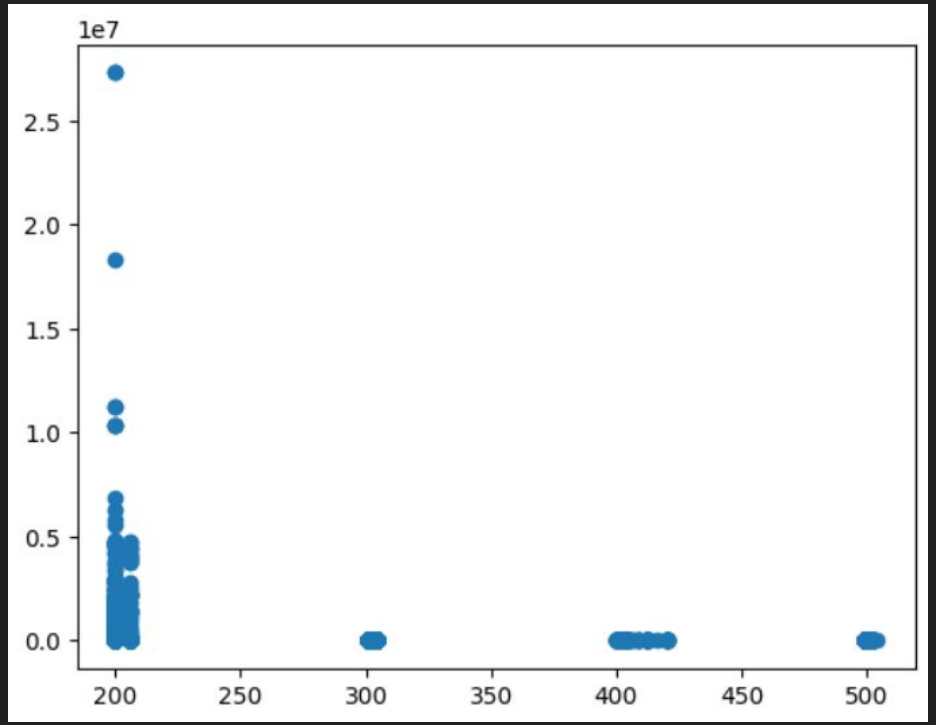
- Entender el uso real de las plataformas (insights)
  - Adaptar a las necesidades reales
  - Toma de decisiones de desarrollo basada en datos
- Escalar apropiadamente la infraestructura
  - Dado las arquitecturas infrastructure as a service (IaaS), es importante tener sólo las prestaciones necesarias
- Detectar comportamiento anómalo
  - Servicios caídos
  - Ataques de denegación de servicio
  - Conexiones anómalas y/o maliciosas

En este trabajo nos centramos en el tercer punto “**Detectar comportamiento anómalo**”

Modelo de Reeb

Se tomaron las variables cantidad de bytes respondidos (server->cliente) y estatus de petición.

Cómo se puede apreciar un comportamiento normal es aquél que un status 200 regresa más bytes.



Modelo de Reeb

Para generar el grafo de Reeb se empleó la **librería “kmapper”** la cuál hace uso del algoritmo DBSCAN para clusterizar los datos.

```
graph = mapper.map(projected_data,data,
                   clusterer=sklearn.cluster.DBSCAN(eps=10000, min_samples=500),
                   cover=km.Cover(n_cubes=10))

# Visualize it
html = mapper.visualize(graph, path_html="test.html",
                 title="Test")
jupyter.display(path_html="test.html")
```

```
Mapping on data shaped (78955, 2) using lens shaped (78955,)

Creating 10 hypercubes.

Created 2 edges and 6 nodes in 0:00:21.050206.
Wrote visualization to: test.html
```

Modelo de Reeb - Parámetros

```
graph = mapper.map(projected_data,data,
                   clusterer=sklearn.cluster.DBSCAN(eps=10000, min_samples=500),
                   cover=km.Cover(n_cubes=10))
```

Recordando los parámetros de DBSCAN:

- eps -> maximum distance between two samples for one to be considered as in the neighborhood
- min_samples -> number of samples (or total weight) in a neighborhood for a point to be considered as a core point

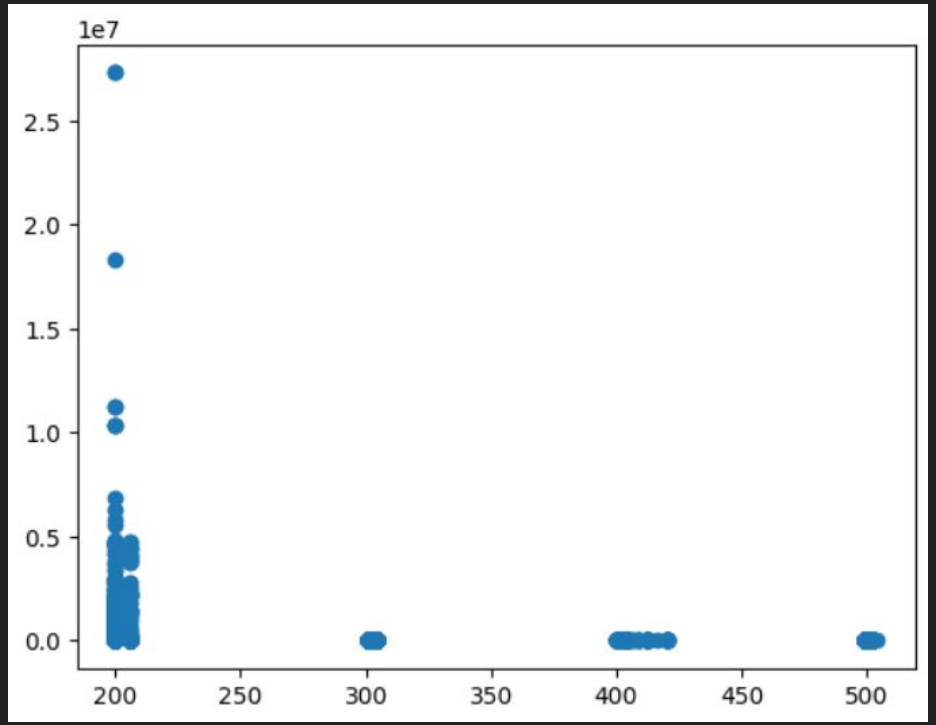
Modelo de Reeb - Parametros

La gráfica cuenta con 78955 puntos (filtrados),

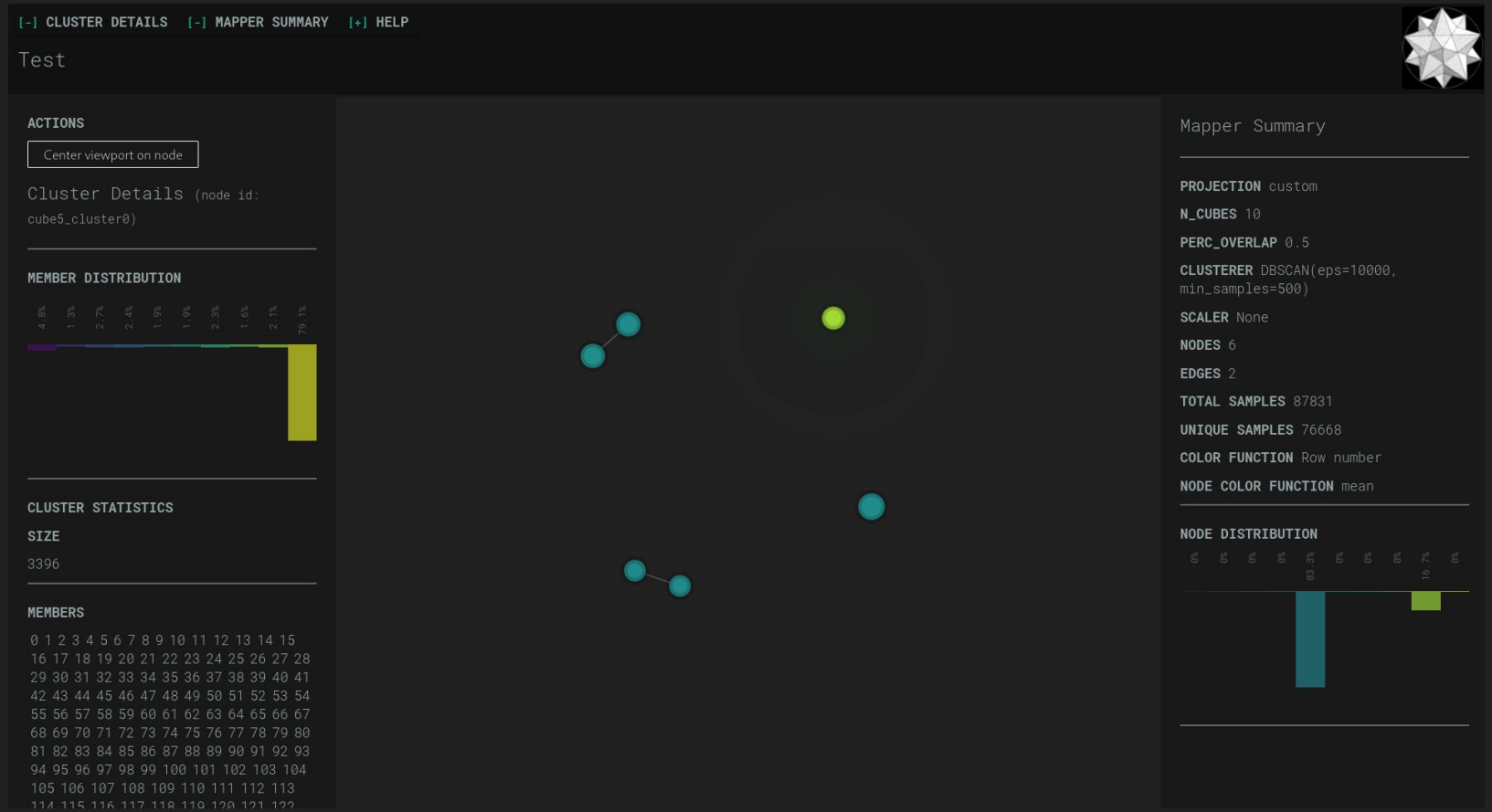
Lo cual nos da un punto de partida para “min_samples”.

Respecto a eps, nuestras unidades son en bytes y en estatus (categóricas) por lo que también debe ser alto

```
eps=10000, min_samples=500
```



Modelo de Reeb



Conclusiones

- Kmapper consume muchos recursos, inicialmente comencé con un muestreo de 60k registros pero consumía todos los recursos
- Creo que jugando más con eps y min_samples es posible empujar más datos y generar gráficos más descriptivos y detallados
- Un problema que tuve es al intentar ingresar datos de fecha (datetime) kmapper no los procesaba
- Con un modelo de Reeb acompañado de distancia de Wasserstein se puede comenzar a detectar comportamiento anómalo