Clustering

Consulting Project

Python and Spark

- You're becoming world famous due to your machine learning skills!
- A technology start-up in California needs your help!

Python and Spark

It's time for you to go to San Francisco to help out a tech startup!

Python and Spark

They've been recently hacked and need your help finding out about the hackers!

Python and Spark

Luckily their forensic engineers have grabbed valuable data about the hacks, including information like session time, locations, wpm typing speed, etc.

Python and Spark

- The forensic engineer relates to you what she has been able to figure out so far, she has been able to grab meta-data of each session that the hackers used to connect to their servers.
- These are the features of the data.

Python and Spark

- `'Session_Connection_Time’` : How long the session lasted in minutes
- `'Bytes Transferred’` : Number of MB transferred during session
- `'Kali_Trace_Used’` : Indicates if the hacker was using Kali Linux
- `'Servers_Corrupted’` : Number of server corrupted during the attack
- `'Pages_Corrupted’` : Number of pages illegally accessed
- `'Location’` : Location attack came from (Probably useless because the hackers used VPNs)
- `'WPM_Typing_Speed’` : Their estimated typing speed based on session logs.

Python and Spark

- The technology firm has 3 potential hackers that perpetrated the attack.
- They are certain of the first two hackers but they aren't very sure if the third hacker was involved or not.
- They have requested your help!

Python and Spark

- Can you help figure out whether or not the third suspect had anything to do with the attacks, or was it just two hackers?
- It's probably not possible to know for sure, but maybe what you've just learned about Clustering can help!

Python and Spark

- One last key fact, the forensic engineer knows that the hackers trade off attacks.
- Meaning they should each have roughly the same amount of attacks.
- For example if there were 100 total attacks, then in a 2 hacker situation each should have about 50 hacks, in a three hacker situation each would have about 33 hacks.

Python and Spark

The engineer believes this is the key element to solving this, but doesn't know how to distinguish this unlabeled data into groups of hackers.

Python and Spark

- Best of luck with this project, it should be a fun one!
- Enjoy!