Microsoft: DAT210x Programming with Python for Data Science

Bookmarks

5. Data Modeling > Lab: Clustering > Assignment 2

Bookmark

Lab Assignment 2

The spirit of data science includes exploration, traversing the unknown, and applying a deep understanding of the challenge you're facing. In an academic setting, it's hard to duplicate these tasks, but this lab will attempt to take a few steps away from the traditional, textbook, "plug the equation in" pattern, so you can get a taste of what analyzing data in the real world is all about.

After the September 11 attacks, a series of secret regulations, laws, and processes were enacted, perhaps to better protect the citizens of the United States. These processes continued through president Bush's term and were renewed and and strengthened during the Obama administration. Then, on May 24, 2006, the United States Foreign Intelligence Surveillance Court (FISC) made a fundamental shift in its approach to Section 215 of the Patriot Act, permitting the FBI to compel production of "business records" relevant to terrorism investigations, which are shared with the NSA. The court now defined as *business records* the entirety of a telephone company's call database, also known as Call Detail Records (CDR or *metadata)*.

News of this came to public light after an ex-NSA contractor leaked the information, and a few more questions were raised when it was further discovered that not just the call records of suspected terrorists were being collected in bulk... but perhaps the entirety of Americans as a whole. After all, if you know someone who knows someone *who knows someone*, your private records are relevant to a terrorism investigation. The white house quickly reassured the public in a press release that "Nobody is listening to your telephone calls," since, "that's not what this program is about." The public was greatly relieved.

Lab

Lecture: Regression
Quiz

Lab: Regression
Lab

Dive Deeper

The questions you'll be exploring in this lab assignment using K-Means are: exactly how useful **is** telephone metadata? It must have some use, otherwise the government wouldn't have invested however many millions they did into it secretly collecting it from phone carriers. Also what kind of intelligence can you extract from CDR metadata besides its face value?

You will be using a sample CDR dataset generated for 10 people living in the Dallas, Texas metroplex area. Your task will be to attempt to do what many researchers have already successfully done - partly de-anonymize the CDR data. People generally behave in predictable manners, moving from home to work with a few errands in between. With enough call data, given a few ***K***-locations of interest, K-Means should be able to isolate rather easily the geolocations where a person spends the most of their time.

Note: to safeguard from doxing people, the CDR dataset you'll be using for this assignment was *generated* using the tools available in the Dive Deeper section. CDRs are at least supposed to be protected by privacy laws, and are the basis for proprietary revenue calculations. In reality, there are quite a few public CDRs out there. Much information can be discerned from them such as social networks, criminal acts, and believe it or not, even the spread of decreases as was demonstrated by Flowminder Foundation paper on Ebola.

1. Open up the starter code in /Module5/**assignment2.py** and *read* through it all. It's long, so make sure you understand everything that is being asked for you before proceeding.
2. Load up the CDR dataset from /Module5/Datasets/**CDR.csv**. Do your due diligence to make sure it's been loaded correctly and all the features and rows match up.
3. Pick the first unique user in the list to examine. Follow the steps in the assignment file to approximate where the user lives.
4. Once you have a (**Latitude**, **Longitude**) coordinate pair, drop them into Google Maps. Just do a search for the "{Lat, Lon}". So if your centroid is located at Longitude = **-96.949246** and Latitude = **32.953856**, then do a maps search for "32.953856, -96.949246".

5. Answer the questions below.

Lab Question

(1/1 point)

Use Google Maps to find the location of the following apartment complexes in the Dallas, TX, USA area. Then keeping that information in mind, answer in the question:

Which of these Apartment Complexes does the first user in the CDR likely live at?

○ The Lexington at Valley Ranch

● Spanish Grove Apartments ✓

○ Tenison at White Rock

○ Downtown Dallas Apartments

○ Grand Estates @ Kessler Park

EXPLANATION

Filter your data for just the first user in the CDR data set. Only look at weekends, and make sure your time is < 6am or > 10p. Then use KMeans to find the centroid location of that point.

Once you have the centroid location, enter it into Google Maps, and search for the above apartments around it. The nearest apartment is the correct answer.

You have used 1 of 2 submissions