Assignment04. Origin-Destination Analysis with NetworkX

Due by 11:00 am December 07, 2021

__All__ operations must be performed in Python!

In this assignment, you are asked to implement origin-destination analysis (one of the network analysis methods) using either **New York City Citi Bike data** or **Chicago Transportation Network Providers data** to assess the impact of new transportation systems we have seen arise in the city. You will use your existing knowledge of spatial data manipulation to assess how well these systems serve a greater societal good.

**This assignment does not require developing your own research question or hypothesis. However, if you want to work on your own research rather than a uniform assignment, it would be more than welcome (extra points).*

Assignment Steps:

1. Choose whether you will look at a historic NYC Citibike trip data or Chicago rideshare trip data. You can choose the same data from the Assignment 03 or you can work on the other one.

    - Citi Bike System Data. (Smaller, and less interesting)
      Please choose ONE months period (if you want, you can do a comparative analysis for two months or multiple months).
      Downloadable from https://www.citibikenyc.com/system-data. Click "Downloadable files of Citi Bike trip data" or "Download Citi Bike trip history data".
    - Chicago Transportation Network Providers (TNP) data. (Larger, but most complete!)
      Please choose just ONE day period (if you want, you can do a comparative analysis for two days or multiple days).
      Downloadable from
      https://data.cityofchicago.org/Transportation/Transportation-Network-Providers-Trips/m6dm-c72p
      **It seems like Chicago OpenAPI has a technical issue for some people. For that case, please download the data after filtering through the dashboard instead of using a query.*

2. Explore, munge, and visualize this information for your own reference, and assess the spatial characteristics, **descriptive statistics** (e.g. total bike rides) and attributes of what the data you have chosen to analyze. Use this to help formulate your thinking for next steps.

3. Create a dataframe with node information
    - Nodes would be based on unique start and end locations. Think about what you need to create a node dataframe.
    - For Citi Bike data, you can easily use station information - already standardized.
    - For TNP data, you need to aggregate data points because pick-up and drop-off locations of ride sharing are very random and there is no standardized pick-up/drop-off location across the city. The easiest way to do this is rounding up latitude and longitude. If you round to 3 decimal places, it's approximately 330feet X 330feet grids. If you round to 2 decimal places, data points would be aggregated into approximately 0.6mile X 0.6mile grids. Please choose your granularity and then create new columns using `.round()` function.

4. Create a dataframe with weighted edge (connection) information based on the data points
    - You need to group data by start and end locations.
    - Weighted values mean the number of connections between start and end stations (e.g. 10 Citi Bike rides between station A and station B or 20 taxi rides between X and Y)

5. Create a directed graph (network) using nodes and edges

6. Calculate incoming and outgoing degrees of nodes and identify top 10 locations (stations) with the most incoming and outgoing rides. Please try to interpret your findings.

7. Calculate at least two centrality measures and interpret the locations with the top 10 ranked centrality scores. If you are not familiar with geographic information, try to explore Google maps or any other maps to find any interesting characteristics around the areas with higher centrality scores.

8. Visualize your network
    - If you want to emphasize something, try to use vivid colors or different marker size for readability
    - If you want to show any interesting patterns in the city, please try to zoom in
    - You can bring additional data (e.g. NTA boundaries)

9. Your deliverable will be another Medium blog entry that address the following:
    a. A short write-up on the data you are using

   b. Your OD analysis process
   c. Describe your findings with well-curated maps, tables, and numerical analysis
   d. This should be no more than 1,000 words.

Deliverables

There are two deliverables:

- **"Medium" style post** (**PDF** format)
   a. No more than 1,000 words (excluding visuals and tables)
   b. Include your appendix about additional data (if any) as well.
- **Code** (**Jupyter notebook and PDF format**)
  Please submit your Python code as Jupyter notebook (.ipynb format) with outputs and PDF file.

Metrics for Success

- Completeness of data processing (retrieving node and edge information)
- Completeness of your OD analysis
- Attractiveness and effectiveness of your visualization
- Completeness and clarity of your script
- Clarity of your article

Extra Credits

- Working on your own research questions (**5 point**)
  If you work on your own research question and hypothesis, you will receive **a maximum 5 points** added to your grade for this assignment (depending on the quality of your work). If you are interested in network analysis or OD analysis and would like to dig deeper into data analysis, please feel free to work on your own research questions. If you think you underperformed on the midterm or previous assignments, please try to use this opportunity to make up for your points. Please state **"Research question(s) is developed in this memo"** on the top of your memo.

- Posting your memo to Medium (0.5 point)
  If you actually post your assignment to your personal blog or create a Medium blog post in a non-anonymized way (in that you are willing to put your name to your work and bravely share it publicly), note the course, and pledge to keep it up, you will receive **0.5 point** added to your grade for this assignment. Add the URL at the end of the assignment for evaluation. It should be noted that when you're applying for jobs, people will search for you. This assignment is testament to your analytical chops and therefore, I highly recommend you use this as a way of increasing your digital footprint.