CSCI - 5901 - The Process of Data Science - Summer 2019
Assignment 3

Due date: August 5th, 2019 11:59:59 pm.
The submission must be done through Brightspace.
Teams of 2 students

- Cite any and all resources used.
    - books, websites (other than documentation) like StackOverflow or help from TA's. Be specific.
- I will use plagiarism tools to detect any type of cheating and copying (your code and PDF).
- Write all of your comments and explanations in the code as a text cell.
- Two sides of Data Science (and your mark):
    - Technical // does it work?
        - Quality of code (documentation, naming)
        - Does it work
    - Conceptual // what did you find from the data?
        - Quality of the data insights
        - Quality of the discussion text
    - Questions will be marked individually. Their weights are shown in parentheses after the question number.
- Your submission is (i) collaboration sheet that shows who did which task, (ii) a single Jupyter notebook, and (iii) a PDF (With the compiled results generated by your Jupyter notebook). Filename should be **A3-<your_name1>-<your_name2>.jpynb and A3-<your_name1 >-<your_name2>.pdf**. Please upload to Brightspace. Please include your B# in your Jupyter notebook and PDF.
- **Forgetting to submit these files results in 0 markings for both students.**

In this question, you are going to solve a spatiotemporal problem and work with some streaming techniques. Follow the steps and provide the answer to each question. The total marks in this assignment are 130. If you score more than 100, I will add the extra points for the previous assignments. In case you score 100 in all assignments, I will proportionally add the bonus points to your midterm.

Download Nima Ports datasets from here: Nima_Ports.Zip. This is a shapefile that you are going to load in your notebook. Download AISdata.zip which is a CSV file similar to ferry AIS data in the lab practice. For simplicity, this file has the trajectory generated by only one vessel. If you are eager to explore your solution on multivessel situation, you can merge this file with the AISferry data provided in the lab.

1. Find all the vessels that visited ports in the provided shapefile (Nima_Ports.Zip). For this part, you are going to create a buffer with an appropriate radius around the shape of each all polygons in the shapefile. Second, you are going to find all the AIS messages (from AIS data) that intersect with these ports. (**20 points**)

2. Show the density (i.e., density is the number of AIS messages in a port), of each port on a map by using a colour-coded map. (**20 points**)

3. Now divide the AIS data into data frames with a one-hour interval. Repeat steps 1 and 2 for all of the sub-dataframes. Here each data frame has only information of one hour. Note that if step 1 and 2 you are using the whole AIS data as a one-time interval. In step 3, you are repeating steps 1 and 2 for all of the one-hour intervals. This can generate many plots that you are going to visualize them. You can save all the plots with proper name and title in a folder or generate a matplotlib animation to visualize it. (**20 points**)

4. Select any port you like. Create a temporal chart for the density of messages in that port. Your x is the time and each snapshot of the time has the density of port at a specific hour. (**20 points**)

5. Use concept drift methods on step 4 and find out if there is any drift in the data that can be detected. Try to play with the input parameters and justify the one you chose. Explain why the drift was detected, what characteristics changed? (**25 points**)

6. Cluster the ports based on their message density using DBSCAN and categorize the ports based on traffic (message density). (**25 points**)