Project Proposal NTDS

Data acquisition

Kaggle Dataset called “Cycle Share Dataset” which is the bicycle trip data from Seattle’s cycle share system.

The cycle share system consists of 500 bikes and 54 stations located in Seattle from 2014 to 2016.

This dataset is composed of 3 tables:

- station.csv
- trip.csv
- weather.csv

Then we will need to compute the number of bikes available at each station every hour (this data is not given in the raw data).

Data exploration

- Identify the users (age, gender, frequency of use)
- What is the most popular starting/stop station?
- How does the number of bike available evolve through time for each station?
- How does weather impact bike trips?
- How do bike trip patterns vary by time of the day and the day of the week?
- How do bike trip patterns vary by kind of users?
- Knowing the weather, can we predict the global bike trip patterns (total number of trips in a day for instance)?
- Knowing the weather, can we predict the bike availability at each station with respect to the time of the day?

Data exploitation

Algorithms to be used:

- PCA to analyze correlation between trips and weather for instance
- In order to predict the bike availability through time at a given station: Use baseline classification techniques with the sklearn package
- *Optional*: Build a fully connected graph of stations (weighted by distances between stations) and find a signal to model bike trips. Then we might use a CNN for analyzing this signal.

Evaluation

Criteria of evaluation:

- Extract global trends in our dataset (such as global correlation for instance)
- How often did we predict well? What percentage of accuracy did we reach in the prediction of the bike availability?