NYC CitiBike Data Analysis

Advanced Big Data Analytics

Project Proposal

Bahul Jain - bkj2111
Gaurang Sadekar - gss2147

Introduction

We will be working with CitiBike Data and perform different types of analysis to understand Citi bike usage and trips across different areas of New York City. We will also use this data to build prediction models which give insight into the usage of Citi bikes in relation to external factors.

Goals

1. Create a visualization showing the favorite neighborhoods based on trip information of CitiBike users.
   - Estimating the requirement of bike stations in neighborhoods.
   - Most common rides each month (by station and by neighborhoods)
   - Estimating deficiency or surplus of bikes in neighborhoods based on usage statistics (available bikes and used bikes).
2. Creating a Regression Model to correlate weather (daily average temperature, precipitation, snow depth) impact on CitiBike activity and forecast future usage.

Data

The data we will be working with is publicly available for use at CitiBike Data. CitiBike Trip data includes the following fields:

- Trip Duration (seconds)
- Start Time and Date
- Stop Time and Date
- Start Station Name
- End Station Name
- Station ID
- Station Lat/Long
- Bike ID

- User Type (Customer = 24-hour pass or 7-day pass user; Subscriber = Annual Member)
- Gender (Zero=unknown; 1=male; 2=female)
- Year of Birth

This is data is clean and anonymized for user privacy, and does not include any sensitive information about the riders themselves.

The system data also includes CitiBike membership data, which has rich per day information about memberships, ride numbers and ride durations.

Sources

1. CitiBike Data
2. National Climatic Data Center

Tools & Languages

1. Spark
2. MLlib
3. Language Choice: Python/Scala
4. Statsmodels / Scikit-Learn
5. Google Maps API
6. Google Geocoding API