Comparing bikeshare programs in NYC, D.C. and Chicago

Farahin Choudhury

Bining Lu

Palani Sundaram

Sumeira Taqui

Project objectives

- Looking at bikeshare programs in some of the major US cities:

  New York City (pop: ~ 8.5M), Washington D.C. (pop: ~700K), and Chicago (pop: ~2.7M)
- Doing analysis on bike ridership in each of the cities
- Impact of the weather on bike usage
- Bike ridership prediction by weather parameters
- Bikeshare customer review analysis
  - Yelp
  - TripAdvisor

Tools and languages

- Web-Scraping
  - Splinter, BeautifulSoup
- Python
  - Pandas, numpy, scikit-learn, pyspark, nltk, matplotlib
- Tableau
- Website
  - Html, bootstrap, javascript, leaflet

Linear regression for predicting ridership by weather parameters

- Aggregated trip-count for 2018 by date and merged with corresponding weather data
- Preprocessed non-numeric data features to binary encoded data
- Scaling and normalization of parameters
- Splitting data into training and testing sets
- Fitting the linear regression model (sklearn library) to scaled training data
- Making predictions by scaled test data

Tools used in processing bike reviews

nltk:

- corpus.stopwords
- wordnet
- ngrams
- RegexpTokenizer
- WordNetLemmatizer

wordcloud :

wordcloud

Pyspark

SparkFiles

pymongo

Pandas

Steps followed in processing bike reviews:

Scrapped and binned based on city and review rating
RegexpTokenizer
Stopwords
Lemmatizer using POS Part of speech
Ngrams -- two word analysis
Find frequency of these words
Stored data in mongo
wordcloud for anaysis single image graphs

Project Challenges

- Having enough data for the project (large data sets)
  - Extensive data cleansing and aggregations to generate several smaller aggregated datasets for each analysis metric
- Using machine learning as a major component of the project
  - NLP
  - Encountering inconsistent results

Resources

- Reviews on Divvy (Chicago):
  - TripAdvisor
- Reviews on Citibike (NYC):
  - TripAdvisor
  - Yelp
- Reviews on Capital bike (Washington D.C.):
  - TripAdvisor
  - Yelp
- Weather data:
  - NCDC

- Bikeshare datasets:
  - https://www.bikeshare.com/data/
  - https://www.capitalbikeshare.com/system-data
  - https://www.citibikenyc.com/system-data
  - https://www.divvybikes.com/system-data

We will present our website to show our findings and explain the analysis in further detail:

Bikeshare-Web